Trabajo Final de Grado

**Grado de Ingeniería en Tecnologías Industriales**

# Estudio del rendimiento de técnicas de minería de datos en la predicción de resultados académicos

**MEMORIA**


**Autor:** Alejandro Mascort
**Director:** Luis José Talavera Méndez


**Convocatoria:** Junio 2019




Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona


**Resumen**

El presente documento pretende establecer una primera comparación de diferentes métodos de predicción, la comparación se realiza en base a los resultados que se obtienen con cada método a la hora de predecir si los alumnos de la escuela aprobarán o suspenderán, en su primer intento, las asignaturas pertenecientes al tercer cuatrimestre del grado (Q3). Los métodos que se emplean son la Regresión Logística, los Árboles de Decisión y los *Support Vector Machine* lineales.

Para realizar este proyecto se sigue una famosa metodología empleada en proyectos de esta tipología, la metodología *CRISP-DM*, cada una de las etapas de dicha metodología está adaptada a este trabajo.

La manipulación de datos, creación de modelos y obtención de resultados se realiza través del lenguaje de programación Python, destacando las librerías Pandas y *SciKit-Learn*.

De los análisis de este trabajo se concluye que para la mayoría de las asignaturas del *Q3,* los dos grupos (aprobados y suspensos) no son separables linealmente entre sí y por esta razón los mejores resultados se obtienen de los Árboles de Decisión, a ello se le suma que en la mayoría de las asignaturas el número de aprobados es mucho mayor, y por ello, los resultados favorecen a este grupo. Los suspensos son predichos de forma acertada en más del 50%, únicamente para Electromagnetismo y Mecánica.

En la sección de Trabajos Futuros de este proyecto se muestran algunos mecanismos que permitirían obtener mejores resultados para los métodos empleados y se sugieren posibles nuevos métodos.

# ÍNDICE

# 1. Introducción

Actualmente, el uso de tecnologías de almacenamiento de datos se ha convertido en una herramienta fundamental para las empresas. Esto es debido a que, gracias a una base de datos, existe la posibilidad de generar estrategias para conseguir nuevos clientes o fidelizar a los habituales, se puede ser más competitivo gracias al conocimiento que dichas bases aportan, puede ayudar a mejorar ciertos procesos, ofrecer ayuda en la toma de mejores decisiones, etc. A raíz de la generación masiva de datos, existe una problemática conocida como infoxicación. La infoxicación o intoxicación por información radica en el exceso de información producida, hecho que dificulta la organización, la extracción y la interpretación de ésta de forma efectiva. Aquí es donde entra en juego la minería de datos (*Data Mining*).

La minería de datos hace referencia a las técnicas aplicadas a los datos para conseguir una mejor comprensión de estos a través de la visualización del comportamiento, o de los patrones y tendencias que estos datos pueden seguir. Conocer dichos patrones o tendencias permite facilitar las estrategias a seguir para llegar a alcanzar los objetivos que cualquier empresa o entidad se haya planteado previamente. Obtener un mayor conocimiento sobre los datos otorga una mayor ventaja competitiva.

De forma habitual, la minería de datos se nombra junto a otro concepto e incluso se llega a confundir con el mismo, este es el caso del aprendizaje máquina (*Machine Learning*). El aprendizaje máquina es una disciplina cuyo objetivo es crear sistemas capaces de “aprender” de forma automática, es decir, ser capaz de identificar patrones de muestras de datos, mejorando de forma autónoma con el tiempo sin la necesidad de intervención humana.

El aprendizaje máquina y la minería de datos utilizan los mismos algoritmos para descubrir patrones en los datos, pero la utilidad de cada una es diferente. El aprendizaje máquina está orientado hacia la obtención de resultados y en cambio, la minería de datos está orientada a descubrir conocimiento.

Hasta ahora se ha hablado desde el punto de vista empresarial, pero ese no es el único enfoque que estas técnicas tienen, también puede ser utilizado para la predicción de posibles catástrofes naturales como es el caso de seísmos, aunque la predicción de terremotos es casi inmediata, estas técnicas podrían ampliar el margen de reacción ofreciendo minutos cruciales para evitar resultados tan devastadores. Existen películas y series de temática futurista que se desarrollan en un entorno que parece muy lejano al actual, pero actualmente se están llevando a cabo operaciones que hacen que ese futuro no sea tan distante, como, por ejemplo, la predicción de

posibles crímenes para evitarlos antes de que ocurran, como en el caso de la famosa película de Tom Cruise, *Minority Report*.

## 1.1. Metodología CRISP-DM

En general, llevar a cabo procesos de minería de datos requiere seguir una serie de pasos, de esta manera se consigue abarcar todas las necesidades de un proceso de minería de datos, así como adquirir el enfoque más idóneo para poder llegar a satisfacer los objetivos planteados. Existen múltiples metodologías diferentes, pero, en este proyecto se ha seguido la metodología conocida como *CRISP-DM*.

La metodología *CRISP-DM* de las siglas de *Cross-Industry Standard Process for Data Mining* y hace referencia a un modelo de trabajo que integra seis etapas de un proceso cíclico. Las etapas que integran a dicha metodología son las visualizadas en la figura mostrada a continuación (*ilustración 1*).

**Comprensión del problema**: Incluye los objetivos, conocimiento previo de la situación de partida, los objetivos a alcanzar mediante la minería de datos y el desarrollo de un plan de proyecto.

**Comprensión de los datos o información**: Esta etapa considera inicialmente una metodología para: conseguir datos, realizar una exploración de los datos para conocer la información que dan y la verificación de la calidad de éstos.

**Preparación de los datos**: Una vez se han recogido los datos necesarios, es necesario limpiarlos y adaptarlos a la forma deseada para poder trabajar con ellos posteriormente. La transformación de los datos puede facilitar la identificación de patrones y comportamientos de los datos.

**Modelización**: Esta etapa involucra la manipulación de softwares de minería de datos para el análisis o estudio de los mismos para establecer agrupaciones de aquellos que muestran un comportamiento similar o para la creación de modelos de predicción, entre otros.

**Evaluación**: Esta etapa pretende descubrir el grado de fiabilidad del modelo o modelos obtenidos en la etapa anterior, para así saber cuál de ellos permite satisfacer de forma más completa los objetivos establecidos en las etapas previas.

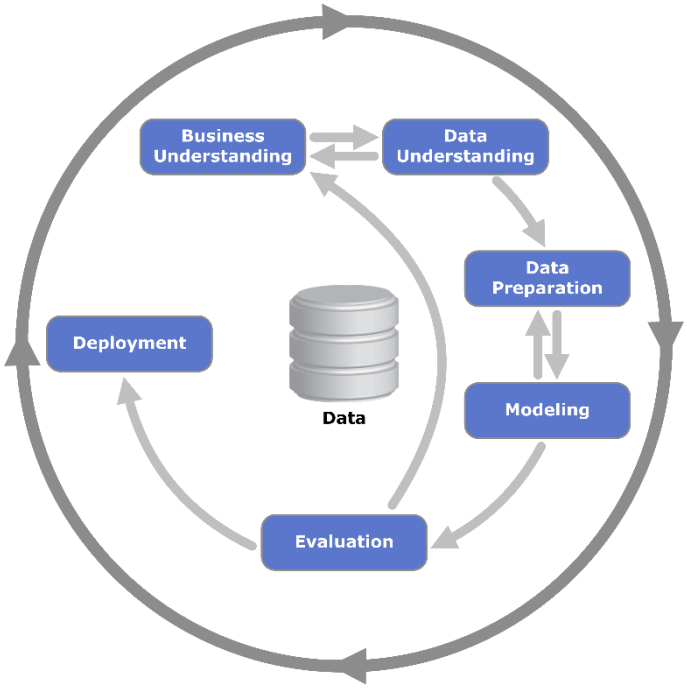


*Ilustración 1. Diagrama de la metodología CRISP-DM.*

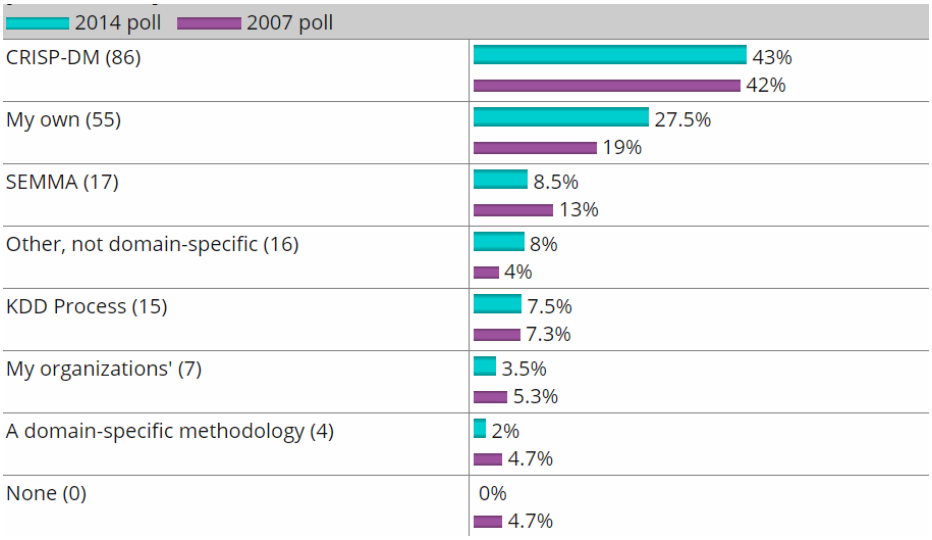


*Ilustración 2. Resultados de las encuestas de los años 2007 y 2014.*

**Implementación**: Es la etapa final de esta metodología, y llega cuando se pretende realizar la implementación del modelo para alcanzar los objetivos establecidos.

La metodología *CRISP-DM* es actualmente la más utilizada para llevar a cabo proyectos de minería de datos. En la *ilustración 2* se puede ver el resultado de una encuesta realizada en el año 2007, y otra en el año 2014, en la prestigiosa web *KDnuggets* (link de los resultados de la encuesta en la bibliografía) sobre la metodología que los usuarios, que llevan a cabo proyectos de *Data Mining*, utilizan.

## 1.2. Objetivos

Este es un proyecto de comparación del rendimiento de métodos de predicción, los diferentes métodos de predicción se aplican sobre un histórico de datos de los alumnos de la ETSEIB (*Escola Tècnica Superior d'Enginyeria Industrial de Barcelona*). Se pretende crear diferentes modelos de predicción con el fin de saber que alumnos aprobarán o que alumnos suspenderán las asignaturas del primer cuatrimestre de la fase no inicial (tercer cuatrimestre del grado). El objetivo principal es el **estudio y la comparación de métodos predicción,** que es conocer el funcionamiento de algunos de los métodos de predicción más empleados, así como entender el porqué de las tasas de acierto en la predicción que obtiene cada uno, y compararlos entre sí, entendiendo porqué funcionan mejor unos que otros en función de los datos. Los métodos de predicción que se estudian son la Regresión Logística, los Árboles de Decisión y el *Support Vector Machine*.

Existen, además del objetivo principal, unos objetivos adicionales necesarios para la realización del proyecto, el primero de ellos es la **aplicación de la metodología *CRISP-DM***, y dentro de esta, la **obtención de conocimiento previo.** También se tiene por objetivo**,** dar una **valoración a los resultados**, ya que es de vital importancia conocer los resultados que se obtienen para poder dar una valoración a cada uno de los modelos, gracias a los resultados se puede comparar el rendimiento de cada método predictivo sobre el conjunto de datos disponible.

Si bien tiene importancia conseguir un máximo de acierto a la hora de realizar predicciones, hay veces en las que se prioriza la obtención de un modelo capaz de predecir de forma correcta un grupo por encima de otro, obteniendo un acierto global inferior. En este caso, se considera que predecir los suspensos tiene mayor relevancia que predecir los aprobados, siempre y cuando el acierto que se obtiene para cada grupo sea correcto. Saber que alumnos van a suspender antes de que ocurra ofrece la posibilidad de evitar que eso suceda, por ejemplo, un alumno que se estima que va a suspender Mecánica, puede recibir algún tipo de ayuda para reforzar su conocimiento y reducir las probabilidades de suspenso, por ello se da énfasis en la **predicción de los suspensos**.

Además de los objetivos anteriores, existen algunos objetivos secundarios del proyecto. Para llevar a cabo cada uno de los objetivos anteriores es necesario disponer de un conocimiento previo que se incrementará a medida que se avance en el desarrollo del proyecto, este conocimiento hace referencia a la familiarización con el entorno de trabajo que se emplea, en este caso *Anaconda*, y la adquisición de una agilidad necesaria para poder manipular los datos

con la librería *Pandas* e implementar cada uno de los métodos de predicción a través de la librería *SciKit-Learn*, ambas librerías pertenecientes a Python.

## 1.3. Abasto del proyecto

Para llevar a cabo la realización del proyecto, se emplea la metodología *CRISP-DM*, en ella se define cada una de las etapas que integran el proyecto de principio a fin. Por ello, a continuación, se muestra cada una de las etapas de la metodología adaptadas a este proyecto.

**Comprensión del problema.** Al ser estudiante de la ETSEIB en el último año de grado, se han cursado todas las asignaturas obligatorias, y por tanto se conoce la exigencia, la dificultad y la metodología de evaluación de cada una de ellas, con ello se procede a la realización del proyecto.

**Comprensión de los datos o información.** Los archivos que se disponen inicialmente contienen los datos de forma ordenada y su comprensión es sencilla. Esta etapa no requiere de gran esfuerzo y se explica posteriormente.

**Preparación de los datos.** A partir de la etapa anterior se manipulan los datos para la obtención de una estructura más ordenada que permita facilitar la implementación posterior de cada uno de los métodos. Cada uno de los pasos que se realizan en esta etapa, se explican en apartados posteriores.

**Modelización.** Es necesario implementar cada uno de los métodos de predicción para la construcción de modelos. La implementación se incorpora a través de Python, algunos métodos admiten de parámetros iniciales para poder construir un modelo.

**Evaluación.** Cada uno de los modelos tendrá que ser verificado sobre datos que no hayan participado en su construcción. Al hacer esta prueba de verificación se puede valorar la fiabilidad del modelo sobre datos “reales”.

**Implementación.** Esta etapa no se incorpora en este proyecto. Para llevar a cabo esta etapa se tendría que poner en funcionamiento cada uno de los modelos en la ETSEIB. Esto se realizaría después de la presentación del proyecto si los resultados fueran satisfactorios e interesara su puesta en marcha.

No todas las etapas que se implementan suponen el mismo esfuerzo, en este proyecto, prácticamente se dedica la mayor parte del tiempo a la realización de las etapas de Preparación, de Modelización y de Evaluación, ya que las restantes o vienen dadas de forma previa o no son necesarias para este proyecto.

# *2.* Comprensión y preparación de los datos

Antes de llevar a cabo la creación de modelos de predicción es interesante conocer como son los datos que se van a emplear. Conocer los datos previamente puede ayudar a entender el porqué de algunos de los resultados que se obtienen.

Los datos que se emplean están organizados en columnas y divididos en diferentes archivos, a pesar de ya disponer de los datos estructurados, para facilitar los análisis a realizar y el empleo de los métodos de predicción, es necesario reorganizarlos y tratarlos para conseguir una estructura aún mejor que facilite las tareas que se vayan a realizar posteriormente.

## 2.1. Comprensión

Inicialmente se dispone de tres archivos Excel correspondientes a diferentes fases de las carreras cursadas en la ETSEIB: fase de preinscripción, fase inicial (primer año de carrera), fase no inicial (segundo año en adelante). Los datos corresponden a los años 2010, en el que se instauró el nuevo programa de la ETSEIB, hasta el año 2017. Cada uno de los archivos fueron proporcionados por el director del proyecto Luís José Talavera.

### 2.1.1. Datos de la fase de preinscripción

Estos datos contienen información de 3709 alumnos (filas) y el archivo tiene las siguientes columnas:

• CODI_EXPEDIENT: Código adjudicado al alumno durante la fase de preinscripción.

• SEXE: Hombre (H) o Mujer (D).

• CP_FAMILIA: Código postal de la residencia familiar del estudiante.

• ANY_ACCES: Año de acceso a la escuela.

• TIPUS_ACCES: Forma de acceso a la escuela (en este caso todas las filas tienen asignado el valor 1).

• NOTA_ACCES: Nota de Selectividad con la que se accede al grado.

• CP_CENTRE_SEC: Código postal del centro de educación secundaria del que proviene.

| CODI_EXPEDIENT | SEXE | CP_FAMILIAR | ANY_ACCES | TIPUS_ACCES | NOTA_ACCES | CP_CENTRE_SEC |
|---|---|---|---|---|---|---|
| 274511 | H | 08640 | 2013 | 1 | 12,99 | 08640 |
| 275156 | H | 43002 | 2013 | 1 | 13,018 | 43002 |
| 259794 | D | 08006 | 2012 | 1 | 12,65 | 08021 |
| 262031 | D | 08017 | 2012 | 1 | 10,74 | 08017 |
| 261879 | D | 08504 | 2012 | 1 | 11,696 | 08500 |
| 258115 | H | 08907 | 2012 | 1 | 10,346 | 08907 |

*Tabla 1. Fragmento de datos del archivo de preinscripción.*

| CODI_PROGRAMA | CODI_EXPEDIENT | CODI_UPC_UD | CREDITS | CURS | QUAD | SUPERA | NOTA_PROF | NOTA_NUM_AVAL | NOTA_NUM_DEF |
|---|---|---|---|---|---|---|---|---|---|
| 752 | 309709 | 240180 | 12 | 2017 | 1 | S | 10 | 10 | 10 |
| 752 | 276774 | 240180 | 12 | 2017 | 1 | S | 8,5 | 8,5 | 8,5 |
| 753 | 229756 | 240232 | 6 | 2011 | 1 | S | 7,5 | 7,5 | 7,5 |
| 753 | 230399 | 240232 | 6 | 2011 | 1 | S | 7,3 | 7,3 | 7,3 |
| 753 | 243272 | 240031 | 6 | 2012 | 1 | S | 7 | 7 | 7 |
| 753 | 247560 | 240231 | 6 | 2011 | 1 | N | 4 | 4 | 4 |
| 753 | 226582 | 240231 | 6 | 2011 | 1 | S | 6,3 | 6,3 | 6,3 |

*Tabla 2. Fragmento de datos del archivo de la fase no inicial.*

### 2.1.2. Datos de la fase inicial y no inicial

Los datos referentes a la fase inicial contienen 53758 filas y los de la fase no inicial, 85736 filas. En las *tablas 1 y 2* se puede ver el formato de los archivos. Las columnas de ambos archivos son las siguientes:

• CODI_EXPEDIENT: Código adjudicado al alumno.

• CODI_UPC_UD: Código de la asignatura a la que corresponde la cualificación.

• CREDITS: Número de créditos de la asignatura.

• CURS: Año en que se cursa la asignatura.

• QUAD: Cuatrimestre en que se cursa la asignatura. (Q1 para el cuatrimestre de otoño i Q2 para el de primavera).

• SUPERA: Aprobado o no aprobado de la asignatura (S = aprobado, N = no aprobado).

• NOTA_PROF: Nota final de la asignatura asignada por el profesor.

• NOTA_NUM_AVAL: Nota final asignada en la evaluación curricular.

• NOTA_NUM_DEF: Nota final definitiva.

## 2.2. Preparación de los datos

A partir de este punto, mediante la librería Pandas de Python, se cargan los tres archivos indicados anteriormente y se ejecutan las acciones correspondientes para facilitar el posterior trabajo con los datos que se disponen. El objetivo es obtener una estructura de datos en la que cada fila corresponda a un alumno. Las operaciones se indican en el orden en el que han sido aplicadas.

### 2.2.1. Filtro y Selección

Primero se filtran los datos de tal forma que los restantes pertenezcan al grado de ingeniería en tecnologías industriales de la ETSEIB, para ello se eliminan los datos con un valor de CODI_PROGRAMA diferente de 752. Una vez aplicado el filtro en los archivos de las fases inicial y no inicial, la columna CODI_PROGRAMA puede ser desechada ya que no aporta ningún tipo información.

En este trabajo, se pretende estudiar el rendimiento de los alumnos de la ETSEIB al cursar por primera vez las asignaturas del tercer cuatrimestre del grado. Esto significa que solo interesan los valores correspondientes a las asignaturas cuyo código es el que se muestra en la *tabla 3*.

Para el archivo que involucra las notas del tercer cuatrimestre, se eliminan aquellas filas que no tengan la variable CODI_UPC con alguno de los valores mostrados en la tabla anterior.

| Código | Asignatura |
|---|---|
| **240132** | Informática |
| **240133** | Mecánica |
| **240131** | Ecuaciones Diferenciales |
| **240033** | Materiales |
| **240031** | Electromagnetismo |
| **240032** | Métodos Numéricos |

*Tabla 3.Tabla de códigos de las asignaturas.*

El siguiente paso a realizar es la ordenación de los datos de las asignaturas del tercer cuatrimestre (datos de la fase no inicial ya filtrados por asignatura). La ordenación se realiza primero por año, luego por cuatrimestre y finalmente, por el código de las asignaturas. Una vez realizada dicha redistribución, aquellos expedientes que se repiten para una misma asignatura se eliminan, es decir, aquellos correspondientes a nuevos intentos para las asignaturas de este cuatrimestre (Q3), ya que se pretende predecir el resultado obtenido la primera vez que se cursa cada una de las asignaturas.

### 2.2.2. Pivotaje

En el siguiente paso, se realiza la construcción de una nueva estructura de datos a partir de los archivos proporcionados, para ello se utilizará el método *pivot_table* de la librería *Pandas* de Python. Este método permite convertir los valores de la variable CODI_UPC_UD en columnas, de modo que cada columna ahora se corresponde con una asignatura y el valor de cada una es la nota obtenida de la asignatura en cuestión. En esta nueva estructura cada fila corresponde a un alumno. Este método se aplica a los archivos de las fases inicial y no inicial. En el caso de que un alumno tenga dos o más notas para una misma asignatura de la fase inicial, el valor que adoptará la nota será la media. Las notas se han tabulado en función de las que indica la variable NOTA_NUM_DEF que es la nota final que consta en el expediente del alumno. En la *tabla 4* se muestra el resultado que se obtiene y la *ilustración 3* indica de forma visual lo que hace el método *pivot_table.*

| CODI_UPC_UD<br>CODI_EXPEDIENT | 240011 | 240012 | 240013 | 240014 | 240015 | 240021 | 240022 | 240023 | 240024 | 240025 |
|---|---|---|---|---|---|---|---|---|---|---|
| **226410** | 6.600000 | 7.6 | 6.30 | 4.15 | 8.000000 | 7.6 | 7.0 | 6.80 | 7.00 | 5.1 |
| **226431** | 5.100000 | 6.0 | 5.60 | 5.30 | 5.600000 | 5.0 | 6.0 | 4.90 | 7.10 | 8.0 |
| **226441** | 3.000000 | 5.0 | 3.65 | 4.00 | 2.500000 | NaN | 0.0 | 3.60 | NaN | NaN |
| **226455** | 8.100000 | 7.2 | 5.80 | 5.00 | 9.200000 | 6.6 | 7.3 | 8.10 | 8.10 | 5.4 |
| **226463** | 3.733333 | 5.2 | 3.55 | 4.80 | 3.233333 | 4.4 | 5.0 | 4.25 | 3.65 | 6.7 |

*Tabla 4. Tabla resultante del pivotaje.*

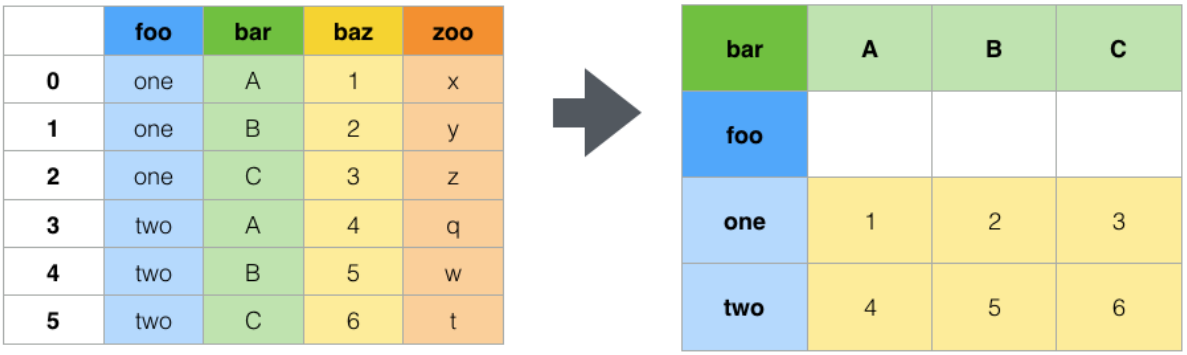


*Ilustración 3. Ejemplo del método pivot_table*

### 2.2.3. Creación de nuevas variables

Seguidamente, se creará la variable *número de intentos* para cada asignatura de la fase inicial, siendo "1" el valor mínimo, es decir, que se ha aprobado con solo un intento. Esta nueva estructura contiene por columnas las asignaturas y el número de intentos de las mismas hasta el año 2017. Si el valor del número intentos para alguna asignatura es *inf* (infinito) quiere decir que para ese momento no se había llegado a superar la asignatura, para estos casos, se eliminará la fila ya que se quiere estudiar el rendimiento de los alumnos que han superado el primer curso completamente, ya que, si no se han superado todas las asignaturas de primer curso, no se pueden haber cursado todas las del tercer cuatrimestre. Además, tener valores de *inf* dificulta la obtención de un modelo predicción.

### 2.2.4. Unión de datos entre archivos

Una vez realizado todo lo anterior, el siguiente paso a realizar es la unión de los archivos de las fases inicial y no inicial (los que han sido modificados hasta tener por columnas las asignaturas y el número de intentos de cada una). La unión se realiza mediante el método *Left_Join* de la librería *Pandas*. Este método complementa al archivo que se especifique (en

este caso el de la izquierda, por eso se llama *left_join*) añadiendo las columnas con sus respectivos valores del archivo restante, este método solo añade a la derecha los valores de las columnas para aquellos CODI_EXPEDIENT que coincidan. El archivo que se sitúa a la izquierda es correspondiente a la fase no inicial y a la derecha el de la fase inicial, de esta forma, solo se añadirán aquellos alumnos que hayan superado asignaturas de la fase inicial.
En este paso, se añade, mediante otro método *Left Join*, la nota de selectividad. A la izquierda estará el archivo anterior y a la derecha estarán las notas de selectividad presentes en el archivo de la fase de preinscripción.

### 2.2.5. Limpieza

Cuando se llevan a cabo operaciones de almacenamiento de datos o de reestructuración de estos, suele perderse información de forma que existen celdas de la tabla que no tienen valor, en este caso, en las celdas aparece el valor *NaN*.
Hasta ahora, se han ido arrastrando valores *NaN* de los archivos iniciales y seguramente haya aparecido alguno a causa de las diferentes transformaciones que los datos han sufrido hasta ahora (mediante el método *Pivot_table).* Los valores *NaN* ofrecen problemas y dificultan los análisis. En este caso las filas que presentan valores *NaN* son una minoría, por tanto, se eliminan sin que ello tenga una repercusión relevante en los análisis posteriores.

### 2.2.6. Transformación y adaptación

Como se pretende crear modelos de predicción para la determinación del aprobado o suspenso de los alumnos, se ha de cambiar el formato de las variables respuesta (asignaturas del tercer cuatrimestre) a formato binario (1 si la nota definitiva está aprobada y 0 si está suspendida).
Por último, se cambiarán los nombres de las columnas. Este paso facilita la interpretación de las columnas. Los nombres de las columnas de la estructura de datos final son los siguientes:

- CODI_EXPEDIENT: código del expediente respectivo a cada alumno.
- ELECTROMAG: nota definitiva de Electromagnetismo
- METODOS: nota definitiva de Métodos Numéricos.
- MATERIALES: nota definitiva de Materiales.
- EDOS: nota definitiva de Ecuaciones Diferenciales.
- INFO: nota definitiva de Informática.
- MEC: nota definitiva de Mecánica.
- ALGEBRA: nota definitiva de Algebra.

- CALCULO1: nota definitiva de Cálculo 1.
- MECFON: nota definitiva de Mecánica Fundamental.
- QUIMICA1: nota definitiva de Química 1.
- FONINFO: nota definitiva de Informática Fundamental.
- GEOMETRIA: nota definitiva de Geometría.
- CALCULO2: nota definitiva de Cálculo 2.
- TERMO: nota definitiva de Termodinámica Fundamental.
- QUIMICA2: nota definitiva de Química 2.
- EXPRE: nota definitiva de Expresión Gráfica.
- Intentos ALGEBRA: número de intentos para aprobar Algebra.
- Intentos CALCULO1: número de intentos para aprobar Cálculo 1.
- Intentos MECFON: número de intentos para aprobar Mecánica Fundamental.
- Intentos QUIMICA1: número de intentos para aprobar Química 1.
- Intentos FONINFO: número de intentos para aprobar Fundamentos de Informática.
- Intentos GEOMETRIA: número de intentos para aprobar Geometría.
- Intentos CALCULO2: número de intentos para aprobar Cálculo 2.
- Intentos TERMO: número de intentos para aprobar Termodinámica.
- Intentos QUIMICA2: número de intentos para aprobar Química 2.
- Intentos EXPRE: número de intentos para aprobar Expresión Gráfica.
- NOTA_ACCES: nota de Selectividad obtenida para acceder al grado.

# 3. Validación

Cuando se trata de modelos de predicción, una de las etapas más importantes es la etapa de validación. Esta etapa permite otorgar una validez el modelo creado, es decir, permite valorar la fiabilidad de cada modelo y realizar comparaciones para escoger aquel que sea capaz de satisfacer los objetivos propuestos en mayor grado. Los múltiples conceptos de relevancia de la etapa de validación se exponen en los siguientes subapartados.

## *3.1.* Conjunto de entrenamiento y *testing*

A la hora de crear un modelo, este se dividirá en dos conjuntos diferentes: uno de entrenamiento y uno de *testing*. El conjunto de datos de entrenamiento se empleará para la creación de los distintos modelos de predicción en función del método utilizado. En cambio, para la verificación de la fiabilidad del modelo, el conjunto será el de *testing*, es decir, para evaluar que tan bien es capaz de predecir dicho modelo a partir de nuevos datos. Si se utilizara el mismo conjunto para entrenar y validar, se estaría haciendo "trampas" ya que se estaría tratando de predecir un resultado que ya ha sido utilizado previamente para la construcción del modelo. De esta forma, es posible que los resultados obtenidos en la predicción fueran satisfactorios y a la hora de utilizar este modelo sobre casos reales, es decir, llevar a cabo predicciones sobre datos en los que aún no se conoce la respuesta, las predicciones fueran nefastas. De esta forma, dividir el conjunto de datos en los anteriores comentados, permite saber la eficacia del modelo antes de ser aplicada sobre casos reales.

## 3.2. Overfitting

El término *overfitting* o sobreajuste se utiliza cuando el modelo se ajusta demasiado bien a los datos del conjunto de entrenamiento y a la hora de aplicarse sobre otros datos nuevos (conjunto de *testing*), este no se ajusta de forma correcta. Lo que ocurre en este caso es, que el modelo solo está aprendiendo de los casos particulares del conjunto de entrenamiento y es incapaz de generalizar para los nuevos. Para evitar el *overfitting* se debe intentar no construir un modelo excesivamente específico. Para ello hay modelos que incluyen parámetros de ajuste como, por ejemplo, la profundidad máxima en los árboles de decisión.

Cuando se construye un modelo sobre un conjunto de datos, y se evalúa sobre el mismo, el resultado que se obtiene puede ser perfecto, es decir, todo lo predicho se acierta. Pero como se

ha indicado en el apartado anterior, modelar y evaluar sobre un mismo conjunto es tratar de predecir algo que ya se conoce, de forma que, si el modelo con brillantes resultados se aplica sobre un nuevo conjunto de datos desconocido para el modelo, existe la posibilidad de que los nuevos resultados sean malos, este sería entonces un claro caso de *overfitting*. Por tanto, para evitar que se produzca *overfitting*, se dividen los datos en conjuntos de entrenamiento y de *testing.*

## 3.3. Validación cruzada

La validación cruzada es una técnica que se utiliza para la evaluación de modelos de minería de datos. Esta técnica permite obtener diferentes modelos de un mismo conjunto de datos y la validación de estos para determinar el grado de variación de los resultados obtenidos.

Pasos de la validación cruzada:

1. Se divide el conjunto de datos en K subconjuntos diferentes.
2. Se combinan K-1 conjuntos y de esta combinación se crea un modelo, el conjunto que no se utiliza en la construcción del modelo, se utiliza en su validación (*testing*).
3. Se repite el paso 2 para una nueva combinación, así hasta haber validado las K combinaciones.

La validación cruzada suele utilizarse sobre el conjunto de *training* (o de entrenamiento) para definir los parámetros más idóneos según el método de predicción que se utilice. Una vez ya han sido definidos, se evalúa el modelo con el conjunto de *testing*.

Existen diferentes metodologías de división de los datos en conjunto de entrenamiento y de *testing*, ahora se verán algunas de las más usadas para llevar a cabo la construcción y validación de los modelos.

### 3.3.1. Validación cruzada aleatoria

Este método consiste en dividir el total de datos de forma aleatoria en conjunto de entrenamiento y de *testing*. Para cada división aleatoria se crea un modelo y se evalúa sobre el conjunto de *testing* resultante. La problemática de esta metodología es que pueden repetirse datos en el conjunto de *testing* y que no se lleguen a utilizar nunca en la creación de un modelo.

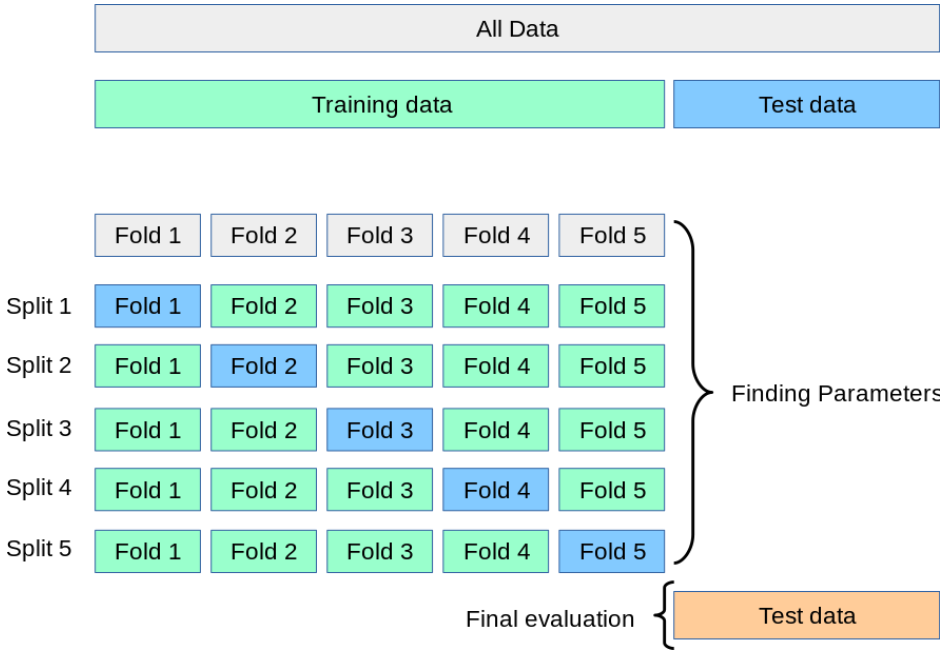


*Ilustración 4. Esquema de Validación Cruzada para 5 iteraciones.*

### 3.3.2. Validación cruzada dejando uno fuera

Este método consiste en dividir el total de datos de forma aleatoria en conjunto de entrenamiento y de *testing*. Para cada división aleatoria se crea un modelo y se evalúa sobre el conjunto de *testing* resultante. La problemática de esta metodología es que pueden repetirse datos en el conjunto de *testing* y que no se lleguen a utilizar nunca en la creación de un modelo.

### 3.3.3. Validación cruzada de K iteraciones

Consiste en dividir los datos en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante K iteraciones, con cada uno de los posibles subconjuntos. Este método es lento desde el punto de vista computacional si se utiliza un elevado número de iteraciones, es por ello por lo que se suele utilizar con 5 o 10 iteraciones. La *ilustración 4* muestra de forma gráfica lo explicado para facilitar la comprensión.

## 3.4. Métricas de evaluación

Dependiendo del objetivo que se tenga, mayor porcentaje posible de acierto al predecir los aprobados, mayor porcentaje posible de acierto al predecir los suspensos u obtener el mayor porcentaje de acierto en la predicción de ambos grupos (aprobado y suspenso), uno se centrará en un término o en otro. Para este proyecto, como ya se ha expuesto en los objetivos, se cree que tratar de determinar que alumnos van a suspender tiene una mayor aplicabilidad. Por ello, el término que más relevancia va a tener será el término *Recall*, explicado a continuación, para la clase 0 (suspenso). Aun así, los otros términos también tienen relevancia ya que estos permitirán valorar si el modelo ha "sacrificado" un elevado porcentaje de acierto en los aprobados para la obtención de una mejora en la predicción del suspenso, y si el precio pagado por dicha mejora merece o no la pena. Por esta razón, se aclara el significado de los siguientes términos utilizados en este trabajo y obtenidos a partir del conjunto de *testing*: Matriz de confusión, *Accuracy Score*, *Precision*, *Recall*, *F1-Score*, *y Support*.

### 3.4.1. Matriz de confusión

La matriz de confusión permite visualizar las clasificaciones que el modelo ha realizado en función de los resultados reales, los términos en la diagonal de dicha matriz/tabla hacen referencia al número de aciertos y los elementos fuera de la diagonal corresponden a los fallos. La estructura de las matrices de confusión se muestra en la *tabla 5*.

### 3.4.2. Tasa de acierto o *Accuracy Score*

El término *Accuracy Score* hace referencia al grado de acierto del modelo creado, es decir, de todos los datos del conjunto de *testing* cuántos han sido clasificados de la forma correcta. Este término se traduce como porcentaje de acierto. Corresponde a la suma de la diagonal de los términos de la matriz de confusión entre el total de datos del conjunto de *testing*.

### 3.4.3. Precisión

Este término se calcula para cada uno de los grupos a predecir y hace referencia al porcentaje de acierto al clasificar cada uno de los grupos, es decir, si se quiere, por ejemplo, calcular dicho término para la clase Suspenso, se hace de la siguiente forma:

$$Precision = \frac{SuspensosAcertados}{SuspensosPredichos}$$

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | Valores que son 0 y se han predicho como 0 | Valores que son 0 y se han predicho como 1 |
| **Valor Real 1** | Valores que son 1 y se han predicho como 0 | Valores que son 1 y se han predicho como 1 |

*Tabla 5. Estructura de las matrices de confusión.*

### *3.4.4. Recall*

Este término se calcula también para cada uno de los grupos de la variable respuesta e indica el porcentaje de acierto del total de cada clase, es decir, si se quiere, por ejemplo, calcular dicho término para la clase Suspenso, se hace de la siguiente forma:

$$Recall = \frac{SuspensosAcertados}{SuspensosTotales}$$

### *3.4.5. Support*

El término *Support* hace referencia al total de datos de cada grupo en el conjunto de *testing*.

# 4. Métodos Predictivos Empleados

En el mundo de la minería de datos existen múltiples técnicas para la creación de modelos predictivos, cada uno de los numerosos métodos funciona de una determinada manera y tiene sus ventajas y sus desventajas. Los algoritmos o técnicas de minería de datos se clasifican en dos grandes grupos: los algoritmos supervisados y los algoritmos no supervisados.

En los algoritmos de aprendizaje supervisado se dispone de datos clasificados por grupos y el objetivo de estos algoritmos es, dadas unas variables de entrada encontrar una función que consiga obtener una relación entre dichas variables y las de salida para poder llevar a cabo predicciones, que asignen la "etiqueta" correcta a cada uno de los nuevos datos que se obtengan.

En los algoritmos de aprendizaje no supervisado no se dispone de datos clasificados por grupos, y por ello el objetivo de estos algoritmos es tratar de encontrar agrupaciones entre los datos de comportamiento similar o que tienen algún tipo de relación, para llevar a cabo simplificaciones sobre estos y obtener información que puede o no tener relevancia.

En este trabajo, se han empleado algunas de las técnicas más utilizadas de aprendizaje supervisado. Se han escogido tres técnicas que funcionan de una forma muy distinta pero que a través de cada una de ellas se pueden llegar a obtener resultados satisfactorios. Las técnicas empleadas y explicadas a lo largo de este punto son: la Regresión Logística, los Árboles de Decisión y las SVM (*Support Vector Machine*). Se escoge la Regresión Logística porque es un método muy popular y sencillo, se añaden las SVM lineales, que son una aproximación similar (al dividir los datos de forma lineal), para comprobar si buscando la separación con un método distinto a la regresión los resultados difieren. Y finalmente, se seleccionan el Árbol de Decisión porque es un método muy diferente a los anteriores.

Aun así, existen muchos más métodos representativos de aprendizaje supervisado que por una cuestión de tiempo no han podido ser implementadas en este trabajo, y que pasarían a ser trabajos futuros para llevar a cabo en una posible continuación de este proyecto.

Para la explicación de cada uno de los tres métodos se utilizarán ejemplos que ayudarán a facilitar el entendimiento de estos antes de que sean aplicados de forma directa sobre el problema de estudio. También se evitará en la medida de lo posible, adentrarse en las matemáticas que hay detrás de cada uno de los algoritmos para poder centrarse exclusivamente en los resultados obtenidos sobre el conjunto de *testing*.

## 4.1. Regresión Logística Binaria

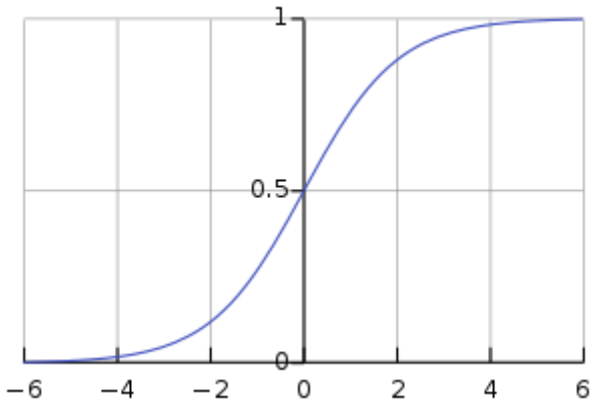


*Ilustración 5. Curva logística.*

La regresión logística corresponde a la suma de los pesos de un conjunto de variables predictoras, pero en vez de ofrecer como resultado directo el valor de dicha suma, ofrece la transformación de dicha respuesta entre los valores 0 y 1 mediante la aplicación de la función logística sobre dicha respuesta. La curva logística se muestra en la *ilustración 5*. Dicho método es utilizado para determinar las probabilidades de que un elemento pertenezca a un determinado grupo. Según este método, si el resultado indica una probabilidad superior al 50 % de pertenecer a un grupo determinado, entonces se clasificará el dato en ese mismo grupo y si es inferior, se clasificará en el restante.

La función de la curva logística es:

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Donde t es el valor de la ecuación de regresión lineal:

$$\mathrm{t}\,(xi) = \theta 0 + \theta 1x1 + \theta 2x2 + \cdots + \theta nxn$$

Siendo:

t: el valor predicho.

n: el número de variables.

$\theta j$: el coeficiente de la variable j-ésima (peso asignado a dicha variable en el modelo).

$x_i$: el valor de la variable para la muestra i (valor de cada una de las variables predictoras en la

muestra i).

Si el valor de $\sigma$ para una muestra es superior o igual a 0.5, el dato se clasifica en el grupo 1 (aprobado) y sino en el grupo 0 (suspenso). El objetivo en este caso es entrenar un modelo de tal forma que se seleccionen los coeficientes $\theta$ más idóneos para obtener el mayor número posible de aciertos. La forma de hacer esto es minimizando una función llamada función de costes.

Como ejemplo, se va a llevar a cabo la regresión logística de la nota obtenida en Mecánica (MEC), mediante la librería de Python, *SciKit-Learn*, en función de la nota de acceso al grado (NOTA_ACCES).

Se ha dividido el conjunto de los datos en un 70% para el conjunto de entrenamiento y en un 30 % para el conjunto de *testing*, seleccionando de forma aleatoria los datos pertenecientes a cada grupo. Luego, se ha procedido a crear un modelo para todo el conjunto de entrenamiento y finalmente, se ha validado el modelo con el conjunto de *testing*.

La recta idónea calculada a partir de los datos del conjunto de entrenamiento es:

MEC = -0.23641510329846935 + 0.020667967327892365*NOTA_ACCES

Según la recta anterior, se puede predecir a partir de qué valor de Nota de Selectividad se clasifican los datos como aprobado. Para conocer el valor, la ecuación a resolver es:

$$0.5 = \frac{1}{1 + \exp(-(-0.23641510329846935 + 0.020667967327892365 * \text{NOTA\_ACCES}))}$$

Obteniendo que el valor a partir del cual se clasificará como 1 será:

NOTA_ACCES = 11.4387

El análisis de los resultados se llevará a cabo más adelante debido a que en este apartado no tiene relevancia.

## 4.2. Árbol de decisión

Es un algoritmo de clasificación supervisado que se utiliza cuando la variable respuesta a predecir es discreta o categórica. Las variables predictoras pueden ser, en cambio, numéricas o categóricas. La ventaja de esta técnica es que la decisión es interpretable de forma gráfica y tiene una fácil comprensión.

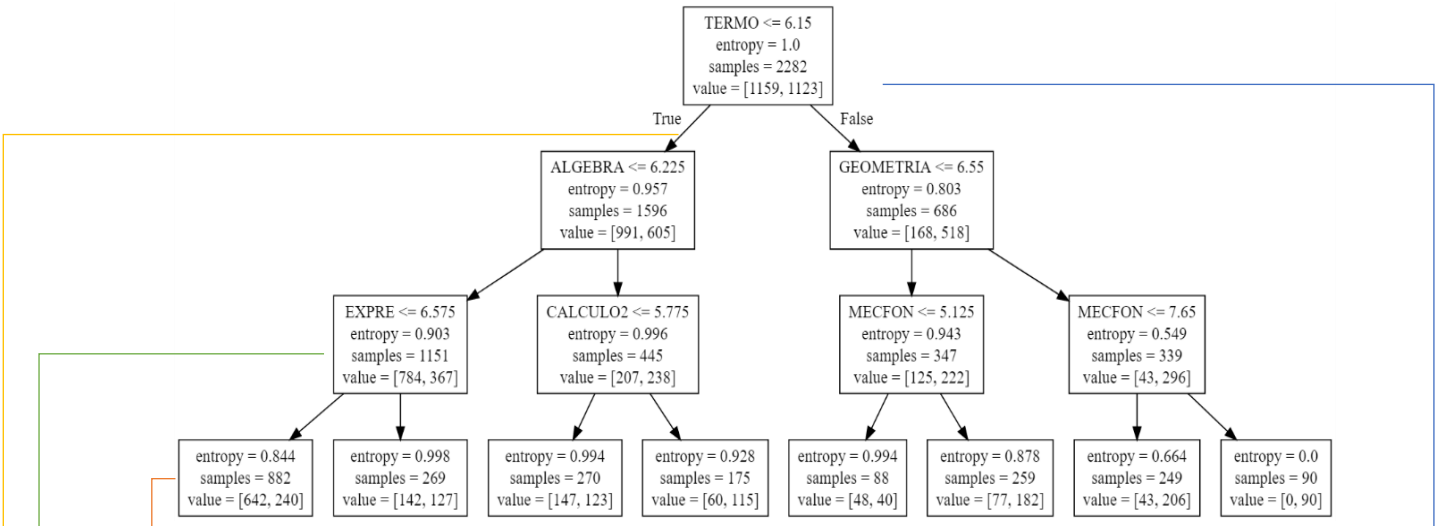


*Ilustración 6. Ejemplo de Árbol de decisión de profundidad 5.*

En este ejemplo, se ha tratado de modelizar la asignatura de Mecánica (MEC), utilizando como herramienta de predicción un árbol de decisión. Para el ejemplo, se han utilizado todas las variables predictoras. En este caso, se ha empleado todos los datos para el cálculo del conjunto de entrenamiento (aquí no tiene relevancia la metodología usada ya que solo se utiliza el ejemplo para visualizar la forma de un árbol de decisión).

Un árbol de decisión contiene:

- ✓ **Nodos:** Representan atributos. Existen diferentes tipos de nodos:
  - **Nodo raíz:** Nodo inicial y a partir del cual se desarrolla el árbol.
  - **Nodo terminal:** Nodo final que indica el resultado de la predicción, si se ha alcanzado dicho nodo.
  - **Nodo interno:** Nodo a partir del cual nace al menos otro nodo.
- ✓ **Rama:** Corresponde a la unión entre dos nodos.

Inicialmente, para el caso del ejemplo, en la *ilustración 6*, se dispone de un total de 2282 datos clasificados, de los cuales, 1159 se clasifican en el grupo 0 (suspenso) y los restantes 1123 en el grupo 1 (aprobado). A medida que se avance en el árbol, el número de datos clasificado se irá reduciendo, provocando una redistribución de estos cada vez más desbalanceada para facilitar la decisión final a la hora de realizar una predicción.

Para recorrer el árbol de decisión y llevar a cabo una predicción, el punto de inicio es el nodo raíz, este nodo pone una condición para escoger el camino a recorrer, en este caso nos indica que si la cualificación correspondiente a termodinámica es inferior o igual a 6.15 el camino a

seguir es el de la izquierda, y si no, el de la derecha. Pongamos que se quiere predecir la nota de un alumno que tiene una nota de 5 en todas las asignaturas, en este caso, todas las condiciones nos obligan a tomar la bifurcación de la izquierda, provocando que el nodo terminal al que se llega sea el de debajo de todo a la izquierda. Si se comprueba la distribución de los datos del conjunto de entrenamiento en este punto, se observa que existen 642 suspensos y 240 aprobados, de esta forma la probabilidad de suspender que se calcula es:

$$\text{Probabilidad de suspenso} = \frac{642}{642+240} \times 100 = 72.79\ \%$$

Y por ello, al ser superior al 50 %, el algoritmo cataloga el nuevo dato como suspenso.

El árbol de decisión mostrado tiene un nivel de profundidad de tres, es decir, el camino más largo hasta el nodo terminal obliga a tomar tres bifurcaciones. En este caso, como se puede ver en la *ilustración 6*, el algoritmo tomaría decisiones con probabilidades cercanas al 50 % en algunos casos, y por ello se estaría cometiendo un elevado grado de error, por esta razón, el nivel de profundidad suele ser muy superior al mostrado.

En la *ilustración 6* aparece otro parámetro llamado *entropy* o entropía. Y es que en realidad, en función del orden de las bifurcaciones, el árbol resultante puede ser uno u otro completamente distinto. Por ello, existe un parámetro que consigue establecer el orden más idóneo. Dicho parámetro es la entropía, ya que sirve para visualizar como de homogéneos son los ejemplos que se quieren clasificar, es decir, la entropía ayuda a medir el grado de incertidumbre del sistema. En este caso, el parámetro utilizado es la entropía, pero existen otros parámetros que permiten la obtención de resultados similares.

Para la entropía:

- Si los datos son completamente homogéneos, es decir, todos pertenecen a una misma clase, el valor es 0.
- Si en cambio tenemos el mismo número de datos para cada clase, el valor es máximo e igual a 1.

El algoritmo, al inicio de creación del árbol, crea diferentes condiciones para cada variable y calcula la entropía resultante para cada una. La condición que haya dado un valor de entropía más bajo o mayor homogeneidad, será la escogida como nodo raíz. Este paso se lleva a cabo de forma iterativa cada vez hasta alcanzar el parámetro de profundidad especificado o hasta haber alcanzado un parámetro de entropía de 0, o nodo completamente homogéneo (también nodo puro). Cuando se alcanza un nodo terminal, ya se puede realizar la clasificación.

## 4.3. Support Vector Machine

El SVM es un algoritmo de aprendizaje supervisado que puede ser utilizado para llevar a cabo clasificaciones lineales y no lineales. Es uno de los modelos más utilizados en *Machine Learning* y es un tipo herramienta adecuada para la clasificación de una cantidad media-baja de datos.

El método que el algoritmo utiliza para llevar a cabo las clasificaciones es mediante la creación de hiperplanos de separación equidistantes de los puntos de cada clase (aprobado o suspenso) más cercanos entre sí. Según la distribución de los datos, se pueden tener datos perfectamente separables y datos que no se pueden separar, en las *ilustraciones 7 y 8* se pueden apreciar estos conceptos de forma gráfica.

Como se puede ver en la *ilustración 7*, existen múltiples rectas que permiten llevar a cabo una división de los datos, por ello existen los márgenes, que son hiperplanos separados de forma equidistante a cada lado del hiperplano de separación y que son calculados únicamente por aquellos puntos que están en la región delimitada por estos, estos puntos también son conocidos como Vectores de Soporte. El objetivo entonces es obtener el hiperplano de separación que consiga la mayor distancia posible entre los márgenes, sin que haya puntos dentro de estos, en el caso de la *ilustración 7*, por tanto, se escogerá como hiperplano de separación el A. También existen funciones que solo son perfectamente separables por funciones no lineales.

En el caso de no tener información separable, como en el caso de la *ilustración 8*, supone una problemática. Para resolver esta problemática, se ha de partir de una premisa base, y es que no todos los datos podrán ser clasificados de forma correcta. En este punto, uno tiene que encontrar el balance que mejor le convenga entre tener los márgenes más anchos posibles y el menor número de puntos dentro dichos márgenes. Para regular este balance, existe el hiperparámetro C, este parámetro creado sirve entonces, para generalizar el error en el modelo.

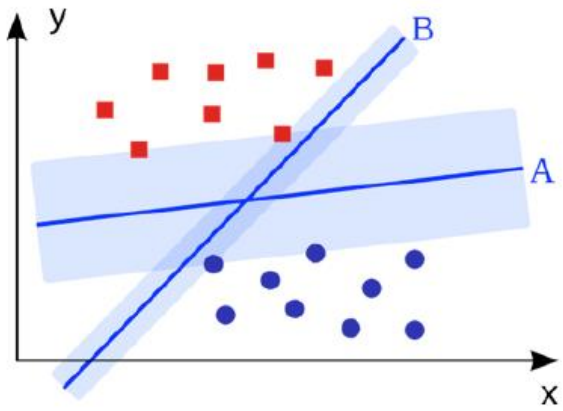


*Ilustración 7. Ejemplo de datos perfectamente Separables.*

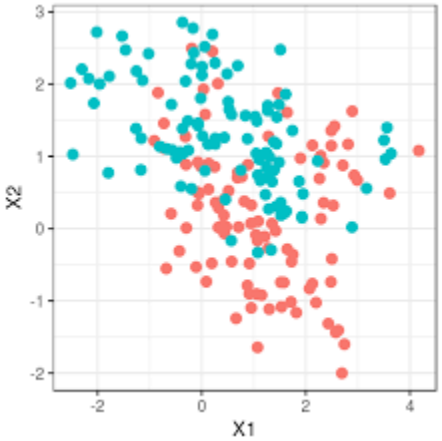


*Ilustración 8. Ejemplo de datos no separables.*

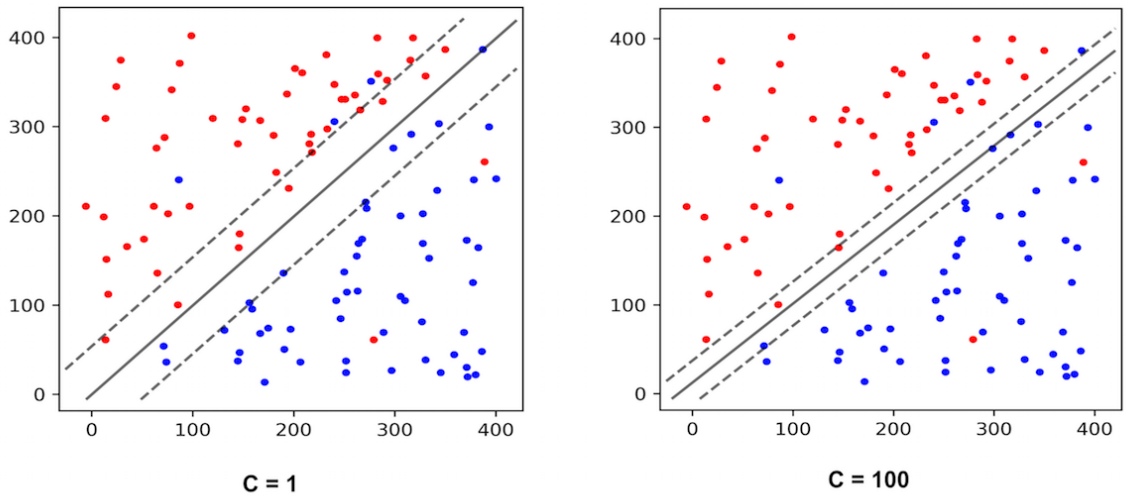


*Ilustración 9. Separación según el parámetro C escogido.*

En la *ilustración 9* se puede visualizar como el parámetro C es capaz de modificar el hiperplano de separación escogido. A mayor valor del parámetro, menor es la distancia entre márgenes, pero menor el número de puntos dentro de estos. Y en cambio, si el parámetro es menor, los márgenes son mayores, y por el contrario la cantidad de puntos dentro de estos también es mayor. En el primer caso, como se puede apreciar, solo 3 azules quedan mal clasificados, es decir, en el lado equivocado del hiperplano de separación. Y en el otro caso, quedarían mal clasificados 4, por lo que respecta al rojo, se cometen 2 errores de clasificación en ambos casos, por ello, en este caso es beneficioso mantener un valor bajo de C, pero esto no será siempre así. En función de los resultados obtenidos en las predicciones, se valorará el parámetro que más convenga para alcanzar los objetivos.

# 5. Resultados

Una vez ya se han introducido los métodos de predicción que se utilizan en este proyecto, se procede a realizar la creación de cada uno de los modelos y a ver los resultados que cada uno de estos ofrece. Antes de empezar se ha de establecer una metodología idéntica para cada modelo que luego permita establecer las comparaciones.

Las variables predictoras utilizadas para cada modelo son las columnas de la estructura de datos final, a excepción de las variables respuesta, se crea un modelo diferente para cada asignatura del tercer cuatrimestre (variables respuesta) y para cada método empleado.

Debido a los costes computacionales de la validación cruzada en este proyecto, se opta por realizar una división de los datos en conjuntos de *training* y *testing*, el primero para crear el modelo y el segundo para validarlo.

Primero de todo, antes de crear cualquier modelo, se dividen los datos en una proporción de 70% para el conjunto de entrenamiento y de 30 % para el *testing*, los valores escogidos no son los únicos posibles, normalmente las proporciones para ambos conjuntos suelen presentar valores en torno al 75% para el training y 25% para el *testing*. La división de los datos se hace de forma aleatoria ya que, al tener los datos en un cierto orden, debido a las transformaciones realizadas para obtener la estructura de datos final, si se dividen de forma que los datos del final pertenezcan al *testing* y los del principio al *training* es posible que los aciertos que se obtengan sean inferiores debido a posibles cambios en las asignaturas a lo largo del tiempo (formato de los exámenes, temario, profesores, exigencia…).

Con los datos divididos, se construyen los modelos y se evalúan. Para los resultados que se obtienen de la validación, se muestra la correspondiente matriz de confusión, la tasa de acierto total, la precisión (*Precision*) del modelo y los aciertos que consiguen para cada grupo (*Recall* de aprobados o suspensos).

## 5.1. Regresión Logística Binaria

Para cada una de las asignaturas se calcula el modelo sin llegar a adentrarse en la verificación de la relevancia de las variables empleadas, esto es debido a que se pretende comparar los diferentes modelos partiendo de unas mismas variables y puede darse el caso en que una variable no sea relevante empleando la regresión logística, pero sea necesaria como nodo para los árboles de decisión. Por ello, es necesario recalcar que, para llegar a obtener modelos

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | 111 | 110 |
| **Valor Real 1** | 76 | 388 |

*Tabla 6.Matriz de confusión para Electromagnetismo*

| **Precisión Suspenso** | **Precisión Aprobado** | **Recall Suspenso** | **Recall Aprobado** | **Tasa Aciertos** |
|---|---|---|---|---|
| 0.59 | 0.78 | 0.50 | 0.84 | 0.73 |

*Tabla 7. Tabla de Precisión, Recalls y Tasa de Aciertos para Electromagnetismo.*

mejores en la regresión logística, es de vital importancia el uso de variables relevantes para el modelo, así como también lo es, la verificación de los residuos (diferencias entre el valor real y el estimado), ya que estos sugieren posibles transformaciones sobre las variables que ayudarían a obtener un modelo mejor. La selección de variables o también conocida como *Feature Selection* es de tal complejidad y extensión que podría dar lugar a otro trabajo entero. Cada variable va acompañada por un coeficiente, y como en este caso, las variables presentan un rango de valores similar, a través de los coeficientes se puede interpretar la relevancia que toma en el modelo cada una de las variables. Siendo la más influyente aquella que presenta el valor más elevado en módulo.

### 5.1.1. Electromagnetismo

La *tabla 6* corresponde a la matriz de confusión que se obtiene al clasificar el conjunto de *testing* con el modelo de regresión obtenido, como se puede ver, el modelo acierta en la clasificación de 111 suspensos y 388 aprobados (términos de la diagonal). Pero comete errores en la clasificación de 186 datos (términos fuera de la diagonal). De los 186 datos, 110 son suspensos que se catalogan como aprobados y 76 son aprobados que se indican como suspensos.

A partir de la matriz de confusión se calculan los términos de la *tabla 7*. El modelo muestra un valor de *Precisión* del 59 % al asignar datos como suspenso y del 78% para aprobados. Si uno se fija en los *Recalls*, puede ver que el porcentaje de acierto es también, mayoritario para los aprobados, se predicen de forma correcta el 84%.

| | Valor predicho 0 | Valor predicho 1 |
|---|---|---|
| **Valor Real 0** | 8 | 83 |
| **Valor Real 1** | 6 | 588 |

*Tabla 8. Matriz de confusión para Métodos Numéricos.*

| Precisión Suspenso | Precisión Aprobado | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|---|
| 0.57 | 0.88 | 0.09 | 0.99 | 0.87 |

*Tabla 9. Tabla de Precisión, Recalls y Tasa de Aciertos para Métodos Numéricos.*

El modelo clasifica correctamente el 50 % de los suspensos. Los resultados obtenidos son los esperados, ya que las dificultades para clasificar los suspensos se pueden deber al desbalance entre aprobados y suspensos, pero también, a que los datos no sean linealmente separables. El grado de acierto total del modelo sobre el conjunto de datos es del 73%, siendo una gran parte debido al acierto en los aprobados.

### 5.1.2. Métodos Numéricos

De la matriz de confusión (*tabla 8)* que se obtiene al clasificar el conjunto de *testing* con el modelo de regresión, resulta que la mayor parte de los aciertos proceden de los aprobados, 588 acertados, por otra parte, se clasifican 8 suspensos de forma adecuada. Los errores en la clasificación proceden, en su mayor parte, de los suspensos predichos como aprobados, un total de 83. De los 89 errores que se cometen al clasificar, en cambio, solo 6 son aprobados que se indican como suspensos.

A partir de la matriz de confusión se calculan los términos de la tabla 9. El modelo muestra un valor de *Precisión* del 57 % al asignar datos como suspenso y del 99 % para los aprobados. Estos datos son engañosos si no se observan los valores referentes a los *Recalls*, ya que visto así el modelo no parece tan malo a la hora de predecir suspensos. Con los *Recalls*, se ve que el porcentaje de acierto es increíblemente fiable para el aprobado, un 99%. Si observamos el *Recall* del suspenso, se puede ver que solo se clasifica de forma adecuada un 9% del total, un valor

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | 66 | 149 |
| **Valor Real 1** | 71 | 399 |

*Tabla 10. Matriz de confusión para Materiales.*

| **Precisión Suspenso** | **Precisión Aprobado** | **Recall Suspenso** | **Recall Aprobado** | **Tasa Aciertos** |
|---|---|---|---|---|
| 0.48 | 0.73 | 0.31 | 0.85 | 0.68 |

*Tabla 11.Tabla de Precisión, Recalls y Tasa de Aciertos para Materiales.*

realmente pobre. El grado de acierto que se obtiene para el conjunto de *testing* es de un 87%. El desbalance exagerado entre suspensos y aprobados en los datos de *testing* (91 suspensos y 594 aprobados) es crítico, causando que los resultados sean tan desbalanceados, a ello se le suma la posible no separación lineal de los datos.

### 5.1.3. Materiales

Para la asignatura de Materiales, como ya se había comentado anteriormente, el balance entre aprobados y suspensos es muy similar al de Electromagnetismo (Materiales: 215 suspensos y 470 aprobados, Electromagnetismo: 221 suspensos y 464 aprobados; en el conjunto de *testing*), por esta razón se comparan ambas asignaturas.

En este caso, la matriz de confusión de la *tabla 10* muestra que el modelo acierta en la clasificación de 66 suspensos y 399 aprobados (términos de la diagonal). Pero comete errores en la clasificación de 207 datos (términos fuera de la diagonal), de estos, 149 son suspensos que se indican como aprobados y 58 son aprobados que se califican como suspensos (para Electromagnetismo son 110 y 76). Estos valores verifican la afirmación anterior, y es que, los datos en Materiales están más mezclados.

El modelo muestra un valor de *Precisión* del 48 % al asignar datos como suspenso y del 73 % para aprobados. Para los *Recalls*, el porcentaje de acierto en los aprobados es del 85 %. En cambio, es muy inferior para los suspensos un 31%. El grado de acierto total para el conjunto de *testing* es de un 68%. El modelo presenta más dificultades en la clasificación de los Suspensos

| | Valor predicho 0 | Valor predicho 1 |
|---|---|---|
| **Valor Real 0** | 8 | 134 |
| **Valor Real 1** | 16 | 527 |

*Tabla 12. Matriz de confusión para EDOS.*

| Precisión Suspenso | Precisión Aprobado | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|---|
| 0.33 | 0.80 | 0.06 | 0.97 | 0.78 |

*Tabla 13.Tabla de Precisión, Recalls y Tasa de Aciertos para EDOS.*

respecto a Electromagnetismo porque seguramente los datos están más repartidos, es decir, puede que existan más alumnos con muy buenas cualificaciones que al cursar por primera vez esta asignatura, la suspendan, y viceversa.

### 5.1.4. EDOS

Para la asignatura de EDOS, los datos están muy desbalanceados, por ello se espera que el grado de acierto para la clase suspenso sea bajo.

De la matriz de confusión (*tabla 12*) que se obtiene al clasificar el conjunto de *testing* con el modelo de regresión, resulta que la mayor parte de los aciertos proceden de los aprobados, 527 bien catalogados, por otra parte, se clasifican 8 suspensos de forma adecuada. Los errores en la clasificación proceden, en su mayor parte, de los suspensos predichos como aprobados, un total de 134. De los 150 errores que se cometen, 16 son aprobados que se indican como suspensos erróneamente.

El modelo muestra un valor de *Precisión* del 33 % al asignar datos como suspenso y del 80 % para los aprobados (*tabla 13*). En la misma tabla, observando los valores referentes a los *Recalls*, se ve que el porcentaje de acierto es muy fiable para el aprobado, un 97%. Si observamos el *Recall* del suspenso, se puede ver que solo se clasifica de forma adecuada un 6% del total, un valor realmente bajo. El grado de acierto de este nuevo modelo para el conjunto de *testing* es de un 78%.

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | 35 | 93 |
| **Valor Real 1** | 38 | 519 |

*Tabla 14. Matriz de confusión para Informática.*

| **Precisión Suspenso** | **Precisión Aprobado** | **Recall Suspenso** | **Recall Aprobado** | **Tasa Aciertos** |
|---|---|---|---|---|
| 0.48 | 0.85 | 0.27 | 0.93 | 0.81 |

*Tabla 15. Tabla de Precisión, Recalls y Tasa de Aciertos para Informática.*

En este caso, el modelo es aún peor que en el de la asignatura de Métodos Numéricos, ya que además de clasificar muy pocos suspensos (24 de 142), no acierta ni en la mitad de ellos.

### 5.1.5. Informática

Para la asignatura de Informática, los datos están muy desbalanceados, por ello se espera que el grado de acierto para la clase suspenso sea bajo. El conjunto de *testing* de Informática (128 suspensos y 557 aprobados) presenta un grado de desbalance similar al de EDOS (142 suspensos y 543 aprobados).

La *tabla 14* corresponde a la matriz de confusión que se obtiene al clasificar el conjunto de *testing* con el modelo de regresión obtenido, como se puede ver, el modelo acierta en la clasificación de 35 suspensos y 519 aprobados (términos de la diagonal). Se comete errores de clasificación en 131 datos (términos fuera de la diagonal). De los 131, 93 son suspensos que se catalogan como aprobados y 38 son aprobados que se indican como suspensos.

De la tabla 15 se observa que el modelo muestra un valor de *Precisión* del 48% al asignar datos como suspenso y del 85% para aprobados. Visualizando los *Recalls*, se puede ver que el valor para los aprobados es de un 93 %. En cambio, el modelo clasifica correctamente el 27 % de los suspensos. El número de aciertos del modelo sobre el conjunto de *testing* es de un 81%, Los resultados obtenidos, a pesar de que distan de ser los ideales, son mucho más satisfactorios que para la asignatura de EDOS, tanto si se mira la *Precisión* como si se mira el *Recall* de los suspensos.

| | Valor predicho 0 | Valor predicho 1 |
|---|---|---|
| **Valor Real 0** | 247 | 76 |
| **Valor Real 1** | 145 | 217 |

*Tabla 16. Matriz de confusión para Mecánica.*

| Precisión Suspenso | Precisión Aprobado | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|---|
| 0.63 | 0.74 | 0.76 | 0.60 | 0.68 |

*Tabla 17. Tabla de Precisión, Recalls y Tasa de Aciertos para Mecánica.*

### 5.1.6. Mecánica

La asignatura de Mecánica muestra un comportamiento diferente a las anteriores, puesto que es la única que esta balanceada e incluso el número de suspensos supera al de los aprobados en el conjunto total de los datos. Para el conjunto de *testing* se tienen 323 suspensos y 362 aprobados.

La *tabla 16* corresponde a la matriz de confusión que se obtiene al clasificar el conjunto de *testing* con el modelo de regresión, como se puede ver, el modelo acierta en la clasificación de 247 suspensos y 217 aprobados. Se comete errores en la clasificación de 221 datos (términos fuera de la diagonal). De los 186 datos, 76 son suspensos que se catalogan como aprobados y 145 son aprobados que se indican como suspensos. En este caso, al contrario que los anteriores, se clasifica la mayoría de los datos como suspenso porque en el conjunto de entrenamiento su mayoría son de ese tipo.

A partir de la matriz de confusión se calculan los términos de la *tabla 17*. El modelo acierta en el 76% de los suspensos y de todos los que asigna acierta el 63%. Este modelo es el que satisface mejor el objetivo de predecir los suspensos. Por lo que respecta a los aprobados el modelo acierta el 60 % del total con un valor de *Precisión* del 74%. Del total, se acierta en el 68% de las clasificaciones. En este caso, el porcentaje de acierto global es inferior al del resto de asignaturas, esto es así porque los valores están repartidos de forma equitativa, ya que en los otros casos el modelo indica casi todos los datos como aprobado y haciendo esto, ya garantiza un porcentaje elevado de acierto.

| | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|
| **Electromagnetismo** | 0.50 | 0.84 | 0.73 |
| **Métodos** | 0.09 | 0.99 | 0.87 |
| **Materiales** | 0.31 | 0.85 | 0.68 |
| **EDOS** | 0.06 | 0.97 | 0.78 |
| **Informática** | 0.27 | 0.93 | 0.81 |
| **Mecánica** | 0.76 | 0.60 | 0.68 |

*Tabla 18. Tabla resumen de los resultados de la regresión logística.*

### 5.1.7. Resumen de la Regresión Logística

A partir de los datos del conjunto de entrenamiento se crean modelos para cada una de las asignaturas, los resultados difieren para cada una. En este caso, el objetivo es clasificar de forma correcta los datos del conjunto de *testing*, dando prioridad a los suspensos.

Es posible que no exista una separación lineal para estos datos, además, como cada una de las asignaturas presenta un desbalance para el número de aprobados y de suspensos, a excepción de Mecánica, los datos referentes a los suspensos (grupo minoritario) son clasificados de forma errónea en su mayoría.

Si se visualizan los valores de los *Recalls* de la *tabla 18* para los suspensos, el modelo que los predice mejor corresponde a la asignatura de Mecánica. En cambio, para las asignaturas desbalanceadas los valores son mucho menores. Las asignaturas en las que se han clasificado peor los suspensos son Métodos Numéricos y EDOS, con valores inferiores a un 10%, y por tanto el modelo creado para cada una no satisface los objetivos planteados de forma satisfactoria. Los suspensos clasificados correctamente para Materiales e Informática presentan también valores bajos, y el único aceptable para del grupo de asignaturas desbalanceadas es Electromagnetismo, que al menos es capaz de acertar en la mitad de los valores clasificados.

Los modelos de las asignaturas con *Recalls* de aprobado más elevados son aquellos que presentan un grado de acierto superior en el total de los de los resultados ya que la mayoría de los datos pertenecen a esa misma clase.

## 5.2. Árboles de decisión

Como ya se ha visto antes, los Árboles de Decisión pueden presentar una elevada profundidad y un número mínimo de datos bajo por nodo, esto permite a los árboles realizar un mayor número de subdivisiones sobre los datos y establecer regiones de decisión más complejas, pero también puede producir un ajuste excesivo sobre los datos provocando un *overfitting*, de forma que al evaluar sobre el *testing* se obtengan malos resultados. Por esta razón, se limitan los valores referentes a los parámetros de profundidad y de mínimo número de datos para formar un nodo.

Para determinar el resultado más favorable en la predicción del suspenso, se combinan valores para ambos parámetros (de 1 a 50), y para cada combinación se visualiza: el grado de acierto total, el grado de acierto del total de suspensos (*Recall* Suspenso) y el grado de acierto del total de aprobados (*Recall* Aprobado). Los valores que se obtienen de cada combinación se muestran en el apartado de *Árboles de Decisión* del *Anexo*. En este apartado se muestran directamente los resultados que se obtienen al haber escogido el modelo que cumple los objetivos en mayor grado para cada una de las asignaturas.

### 5.2.1. Electromagnetismo

En la *tabla 19* se muestra la combinación de valores, profundidad y número mínimo de elementos por nodo, para el árbol que obtiene un *Recall del* Suspenso máximo, así como la tasa de acierto global que obtiene el modelo.

El resultado óptimo de predicción del suspenso se encuentra para una profundidad de 10 y un número mínimo de elementos de 7, con este árbol de decisión se obtiene un acierto global de un 67% sobre el conjunto de *testing*.

La *tabla 20* corresponde a la matriz de confusión de las predicciones que el modelo realiza. El máximo número de suspensos predichos utilizando un árbol de decisión es de 132 suspensos. Se cometen 137 errores al asignar datos como suspenso sin que estos lo sean. Para los aprobados el modelo acierta en 327 y comete el error de asignar como aprobado a 89 que no lo son.

En la *tabla 21* se tabulan los resultados obtenidos referentes a *Recalls* y *Precisión*, el valor de *Recall* de Suspenso máximo que se puede obtener es de un 60%, para este árbol se obtiene un *Recall 1* de un 71%. El modelo pierde en precisión ya que únicamente acierta en el 49% de los casos que asigna como suspensos, para el aprobado el porcentaje es mayor, un 79%.

| Profundidad | Nº Elem. Mínimo | Tasa Aciertos |
|---|---|---|
| 10 | 7 | 0.67 |

*Tabla 19. Tabla de combinación óptima y Score. (Electromagnetismo)*

| | Valor predicho 0 | Valor predicho 1 |
|---|---|---|
| **Valor Real 0** | 132 | 89 |
| **Valor Real 1** | 137 | 327 |

*Tabla 20. Matriz de confusión de Recall máximo para Electromagnetismo.*

| Precisión Suspenso | Precisión Aprobado | Recall Suspenso | Recall Aprobado |
|---|---|---|---|
| 0.49 | 0.79 | 0.60 | 0.71 |

*Tabla 21. Tabla de Precisión y Recalls para Electromagnetismo.*

| Profundidad | Nº Elem. Mínimo | Tasa Aciertos |
|---|---|---|
| 20-30;45 | 10 | 0.82 |

*Tabla 22. Tabla de combinación óptima y Score. (Métodos Numéricos)*

| | Valor predicho 0 | Valor predicho 1 |
|---|---|---|
| **Valor Real 0** | 30 | 61 |
| **Valor Real 1** | 62 | 532 |

*Tabla 23. Matriz de confusión de Recall máximo para Métodos Numéricos.*

| Precisión 0 | Precisión 1 | Recall Suspenso | Recall Aprobado |
|---|---|---|---|
| 0.33 | 0.90 | 0.33 | 0.89 |

*Tabla 24. Tabla de Precisión y Recalls para Métodos Numéricos*

### 5.2.2. Métodos Numéricos

En la *tabla 22* se muestra la combinación de valores, profundidad y número mínimo de elementos por nodo, para el árbol que obtiene un *Recall* del Suspenso máximo, así como el grado de acierto global que obtiene el modelo de predicción para la asignatura de Métodos Numéricos.

Para la asignatura de Métodos Numéricos, existen varias combinaciones que permiten obtener un mismo valor máximo de predicción de los suspensos para el conjunto de *testing*, todas las combinaciones que maximizan este valor tienen en común que el número mínimo de elementos para crear un nodo ha de ser de 10. Las profundidades óptimas en este caso corresponden a los valores: 20, 25, 30 y 45.

De la matriz de confusión de la *tabla 23*, se obtiene que el número de aciertos de los suspensos es de 30, en cambio para los aprobados es muchísimo mayor, 532. Esto es así, porque como ya se ha comentado de forma repetida, en esta asignatura existen muchos más aprobados que suspensos. Se han asignado 62 suspensos y 61 aprobados de forma incorrecta.

Los valores de *Precisión* y *Recall* son muy similares en ambos grupos, para los suspensos son un 33% y 33% respectivamente, para los aprobados son un 90% y un 89%. El modelo presenta facilidad en la predicción de los aprobados y dificultad en la asignación de los suspensos.

### 5.2.3. Materiales

Para la asignatura de Materiales existen tres combinaciones diferentes que maximizan la predicción de los suspensos, entre las cuáles hay de dos tipos: o un número mínimo de elementos para la creación de un nodo muy elevado con una profundidad baja o, un elevado número de profundidad, pero con un número mínimo de elementos para la creación de un nodo muy bajo.

Para cada una de las combinaciones, se visualizan los valores que se obtienen sobre el conjunto de *testing.* Los resultados aparecen en las *tablas 25, 26 y 27*.

Las combinaciones que obtienen el valor máximo de *Recall del Suspenso* son: una profundidad de 7 con un número mínimo de elementos por nodo de 50, una profundidad de 35 con un mínimo de 3 elementos por nodo y una profundidad de 50 con 2 elementos por nodo como mínimo. De todos los aciertos globales, el más elevado es el que corresponde a la primera combinación con un 68% de acierto.

Como el mejor modelo es el de la primera de las combinaciones, ya que el valor de *Recall* del Suspenso para todos ellos es el mismo pero las predicciones del aprobado son mejores, solo se muestra la matriz de confusión que se obtiene con este modelo (*tabla 25*). El modelo acierta 94 suspensos y 371 aprobados. Existen 99 suspensos y 121 aprobados que se asignan de forma errónea.

| Profundidad | Nº Elem. Mínimo | Tasa Aciertos |
|---|---|---|
| 7;35;50 | 50;3;2 | 0.68;0.64;0.63 |

*Tabla 25.Tabla de combinación óptima y Score. (Materiales)*

| | Valor predicho 0 | Valor predicho 1 |
|---|---|---|
| **Valor Real 0** | 94 | 121 |
| **Valor Real 1** | 99 | 371 |

*Tabla 26. Matriz de confusión de Recall máximo para Materiales.*

| Precisión Suspenso | Precisión Aprobado | Recall Suspenso | Recall Aprobado |
|---|---|---|---|
| 0.49 | 0.79 | 0.44 | 0.75 |

*Tabla 27. Tabla de Precisión y Recalls para Materiales.*

En la *tabla 27* se muestran los resultados finales, la precisión del modelo para el suspenso es de un 49% y para el aprobado de un 79%, el *Recall* máximo del suspenso que se puede obtener es un 44% y para este mismo modelo, el *Recall* del aprobado es de un 75%.

### 5.2.4. EDOS

En el caso de EDOS, asignar un valor de profundidad de tres o inferior provoca que el árbol no sea capaz de predecir de forma correcta ningún suspenso. A partir de una profundidad de 5, se aciertan más del 20 % de los suspensos independientemente del valor que se asigne al número mínimo de elementos por nodo.

En las *tablas 28, 29 y 30* se muestra la combinación escogida, el porcentaje de acierto, la matriz de confusión que se obtiene sobre el conjunto de *testing*, los valores de *Precisión* para cada grupo y los *Recalls*.

Escoger una profundidad de 45 y un número mínimo de elementos por nodo de 5 provoca que se obtenga el máximo acierto en el grupo de los suspensos.

El acierto total es de un 74%, se clasifican correctamente un total de 505 datos, 50 suspensos y 455 aprobados. De todos los valores asignados como suspenso, 88 son mal asignados y para los aprobados hay 92 mal asignados.

| Profundidad | Nº Elem. Mínimo | Tasa Aciertos |
|---|---|---|
| 45 | 5 | 0.737 |

*Tabla 28. Tabla de combinación óptima y Score. (EDOS)*

| | Valor predicho 0 | Valor predicho 1 |
|---|---|---|
| **Valor Real 0** | 50 | 92 |
| **Valor Real 1** | 88 | 455 |

*Tabla 29. Matriz de confusión de Recall máximo para EDOS.*

| Precisión Suspenso | Precisión Aprobado | Recall Suspenso | Recall Aprobado |
|---|---|---|---|
| 0.36 | 0.83 | 0.35 | 0.84 |

*Tabla 30. Tabla de Precisión y Recalls para EDOS.*

La precisión del modelo para los suspensos es de un 36%, para los aprobados, es de un 83%. El *Recall* máximo que se puede obtener con las combinaciones realizadas es de un 35%, para este modelo el *Recall* de los aprobados es de un 84%.

### 5.2.5. Informática

Al igual que en la asignatura de EDOS, existen profundidades en las que se predicen mal todos los suspensos, para Informática, esto solo ocurre en una profundidad de 1.

En este caso existen dos respuestas que maximizan el valor del acierto obtenido para el grupo de los suspensos, pero para ambos modelos, el parámetro referente al número de elementos mínimo para formar un nodo es de dos.

En las figuras X1 se muestran las combinaciones y los *Scores* que resultan. Para una profundidad de 30 el acierto global o *Score* es de un 71.5%, y en cambio, para una profundidad de 45 es de un 70.1%. Se escoge como modelo el primero de ellos, por tener un grado de acierto superior.

Para el modelo que se escoge, se muestran los resultados de la matriz de confusión en la figura X. Según la matriz que se obtiene, se aciertan 55 suspensos y 433 aprobados. Por el contrario, se asignan 124 suspensos y 73 aprobados de forma errónea.

| Profundidad | Nº Elem. Mínimo | Tasa Aciertos |
|---|---|---|
| 30;45 | 2 | 0.715;0.701 |

*Tabla 31. Tabla de combinación óptima y Score. (Informática).*

| | Valor predicho 0 | Valor predicho 1 |
|---|---|---|
| **Valor Real 0** | 55 | 73 |
| **Valor Real 1** | 124 | 433 |

*Tabla 32. Matriz de confusión de Recall máximo para Informática.*

| Precisión Suspenso | Precisión Aprobado | Recall Suspenso | Recall Aprobado |
|---|---|---|---|
| 0.31 | 0.88 | 0.43 | 0.78 |

*Tabla 33. Tabla de Precisión y Recalls para Informática.*

El modelo muestra un 31% de precisión en los suspensos en cambio los aprobados presentan un grado de precisión muy superior, un 86%. El acierto máximo para los suspensos es de un 43% y para los aprobados es de un 78%.

### 5.2.6. Mecánica

La asignatura de Mecánica es la única en la que el número de suspensos es mayor al de aprobados y, por tanto, en este caso, se obtienen valores de acierto muy superiores al 50%. De esta forma, el modelo consigue el valor máximo para una profundidad mínima (de 1). Los resultados que se obtienen para esta profundidad son los mismos sea cual sea la agrupación mínima para la formación de un nodo.

Los resultados se muestran en las *tablas 34, 35 y 36*. Se clasifican adecuadamente el 64% de los datos obtenidos. El grado de acierto, se produce en su mayoría en los suspensos, con un valor de un de un 84%. Para los aprobados es de un 46%, un valor muy inferior, producido por que es el grupo minoritario y por la profundidad escogida. Si se escoge un valor de profundidad mayor, el modelo se ajusta mejor a los aprobados. Pero en este caso, se busca maximizar los suspensos acertados. El número de aciertos para los suspensos es de 270, con una precisión del 58%, y para los aprobados la precisión es de un 76%, se aciertan en total 168 aprobados.

El modelo que se genera para satisfacer los objetivos, en este caso, es el árbol de decisión más simple que se podía crear, un árbol que separa los datos en dos subconjuntos.

| Profundidad | Nº Elem. Mínimo | Tasa Aciertos |
|---|---|---|
| 1 | 1-50 | 0.639 |

*Tabla 34. Tabla de combinación óptima y Score. (Mecánica)*

| | Valor predicho 0 | Valor predicho 1 |
|---|---|---|
| **Valor Real 0** | 270 | 53 |
| **Valor Real 1** | 194 | 168 |

*Tabla 35. Matriz de confusión de Recall máximo para Mecánica.*

| Precisión Suspenso | Precisión Aprobado | Recall Suspenso | Recall Aprobado |
|---|---|---|---|
| 0.58 | 0.76 | 0.84 | 0.46 |

*Tabla 36. Tabla de Precisión y Recalls para Mecánica.*

### 5.2.7. Resumen Árboles de decisión

En la *tabla 37* se muestran los resultados para cada uno de los modelos. Aquellos modelos de asignaturas con un mayor grado de desbalance entre aprobados y suspensos requieren de una profundidad mayor, como es el caso de Métodos Numéricos, EDOS e Informática. Esto es debido a que el árbol necesita llevar a cabo un mayor número de subdivisiones para poder identificar el comportamiento de los suspensos. En cambio, para las asignaturas menos desbalanceadas, las profundidades no superan el valor de 10, este es el caso de Electromagnetismo, Materiales y Mecánica.

El parámetro del número de elementos mínimo para la creación de un nodo es bajo para aquellos modelos que tienen una profundidad elevada ya que, a mayor número de subdivisiones, menor número de elementos en cada subconjunto. Al contrario, ocurre con los modelos que presentan baja profundidad, que permiten, en general, un valor de agrupación mínima más elevado.

Los modelos que consiguen los mejores resultados en la clasificación de suspensos son los de Mecánica y Electromagnetismo con unos valores de acierto superiores al 50%. El resto de las asignaturas en cambio, presentan un valor comprendido entre el 30 y el 50 % de acierto. Al contrario, ocurre con los *Recalls* de los aprobados, los valores más elevados se corresponden a las que obtienen *Recalls* de suspenso más bajo. En este caso, Mecánica es la única que no llega a superar el 50 % de aciertos.

| | Recall Suspenso | Recall Aprobado | Tasa Aciertos | Profundidad | Nº Elem. Mínimo |
|---|---|---|---|---|---|
| **Electromagnetismo** | 0.60 | 0.71 | 0.67 | 10 | 7 |
| **Métodos** | 0.33 | 0.89 | 0.82 | 20-30;45 | 10 |
| **Materiales** | 0.44 | 0.75 | 0.68 | 7 | 50 |
| **EDOS** | 0.35 | 0.84 | 0.74 | 45 | 5 |
| **Informática** | 0.43 | 0.78 | 0.72 | 30 | 2 |
| **Mecánica** | 0.84 | 0.46 | 0.64 | 1 | 1-50 |

*Tabla 37. Tabla resumen de los resultados de los árboles de decisión.*

El porcentaje de acierto global es similar al de los aprobados, las que presentan el mayor porcentaje de acierto en los aprobados, al ser el grupo mayoritario (excepto en Mecánica), presentan el mayor porcentaje de acierto global.

## 5.3. SVM

Para crear los modelos de *Support Vector Machine* es necesario asignar un valor al hiperparámetro C, por ello se grafican para cada una de las asignaturas los valores referentes a los *Recalls* de aprobados y suspensos en función del parámetro.

Se escoge un rango de valores que el hiperparámetro C puede tomar, se escoge una amplia variedad: 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000, 1000000, 10000000, 100000000 y 1000000000. Para mantener una distancia regular entre los valores del eje X (valor del hiperparámetro C), se escoge una escala logarítmica para este parámetro a la hora de representarlo.

Se presentan los resultados de cada una de las predicciones en el apartado *Support Vector Machine* del *Anexo*: valores de *Recall*, *Support* y el número de vectores de soporte utilizados para la construcción de cada modelo. Para cada asignatura se escoge el parámetro C que maximiza el acierto en los suspensos, siempre y cuando esto no suponga una penalización elevada en el acierto global.

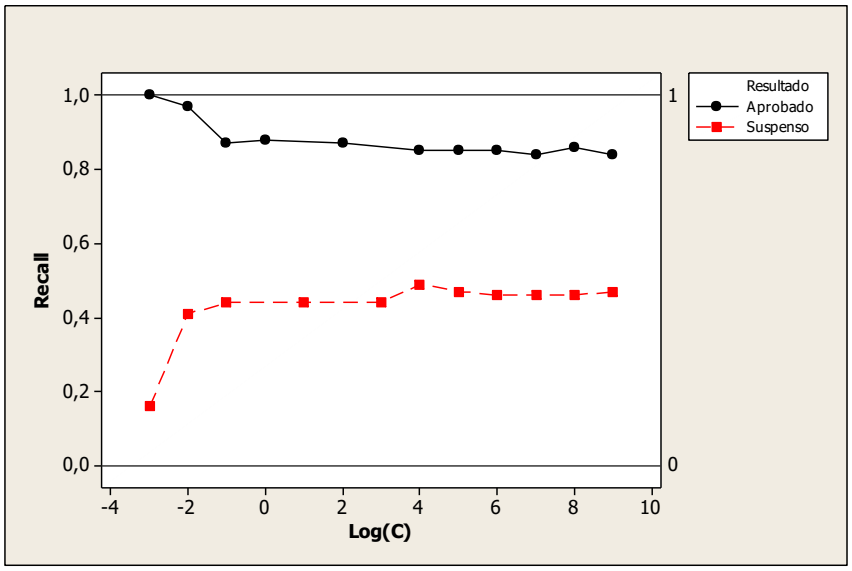


*Ilustración 10.Gráfica de Recalls en función del hiperparámetro C para Electromagnetismo.*

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | 108 | 113 |
| **Valor Real 1** | 70 | 394 |

*Tabla 38. Matriz de confusión de Recall máximo para Electromagnetismo.*

| **C** | **Precisión Susp.** | **Precisión Aprob.** | **Recall Susp.** | **Recall Aprob.** | **Tasa Ac.** |
|---|---|---|---|---|---|
| 10000 | 0.61 | 0.78 | 0.49 | 0.85 | 0.73 |

*Tabla 39. Tabla de resultados de Electromagnetismo. (SVM)*

### 5.3.1. Electromagnetismo

Para la asignatura de Electromagnetismo, si el valor del hiperparámetro C es 0.001, se obtiene un porcentaje de acierto muy bajo en la predicción de suspensos, pero para el resto de los valores, el acierto es mucho mayor y similar, con valores cercanos al 45%.

En la *ilustración 10* se visualiza la evolución de los *Recalls* a medida que se modifica el hiperparámetro. En este caso, aparece un pico máximo para el *Recall* del suspenso cuando C toma un valor de 10000.

El modelo clasifica 108 suspensos de forma correcta, y se equivoca en 70 suspensos que en realidad son aprobados. Para los aprobados acierta 394, y falla en 113 que no lo son.

Para el valor que se escoge, se muestran los resultados en la *tabla 39*. El porcentaje de acierto

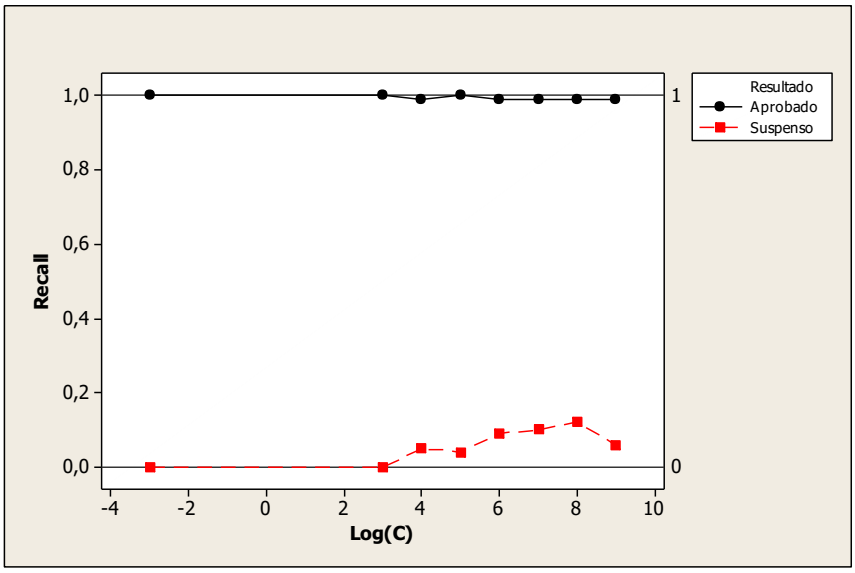


*Ilustración 11.Gráfica de Recalls en función del hiperparámetro C para Métodos Numéricos.*

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | 11 | 80 |
| **Valor Real 1** | 6 | 588 |

*Tabla 40. Matriz de confusión de Recall máximo para Métodos Numéricos.*

| **C** | **Precisión Susp.** | **Precisión Aprob.** | **Recall Susp.** | **Recall Aprob.** | **Tasa Ac.** |
|---|---|---|---|---|---|
| 100000000 | 0.65 | 0.88 | 0.12 | 0.99 | 0.87 |

*Tabla 41. Tabla de resultados de Métodos Numéricos. (SVM)*

para el grupo de los suspensos es de un 49%. De todos los datos que se clasifican como suspenso, el 61% lo es realmente y el 39% restante pertenece datos cuyo grupo real es el de aprobados. El porcentaje de acierto sobre los aprobados es de un 85%. El modelo presenta una precisión del 78% al asignar datos como aprobado. El grado de acierto sobre el total de los datos de *testing* es de un 73%.

### 5.3.2. Métodos Numéricos

El grado de acierto que se obtiene sobre el suspenso con los diferentes valores de C ronda valores muy bajos en todo momento, aun así, como se puede ver en la ilustración 11, existe un máximo para un valor de *Log(C*) de 8. Los *Recalls* de los aprobados, en cambio, presentan en todo momento valores del 100% o muy cercanos a este.

La matriz de confusión de la *tabla 40* muestra que el modelo clasifica 11 de 221 suspensos

correctamente, en los aprobados clasifica 588 de 594 ofreciendo un valor alto de fiabilidad. Existen 80 suspensos mal catalogados, y 6 aprobados también clasificados de forma equivocada.

En la *tabla 41* se muestran los valores que el modelo obtiene con el valor de *Log(C)* de 8. El modelo obtiene un porcentaje de acierto del 12% sobre los suspensos. La precisión con la que el modelo asigna suspensos es de un 65%. En cambio, para los aprobados los valores obtenidos son mucho mayores, se clasifican correctamente el 99% de los valores y se hace con una precisión del 88%. El porcentaje de acierto total es de un 87%.

En esta asignatura los datos están muy desbalanceados, y la mayoría de los que se han utilizado para construir el hiperplano y sus márgenes pertenecen a la clase aprobado, por esta razón, la mayoría son clasificados como tal y el número de aciertos de esta clase es tan elevado respecto a la clase suspenso.

### 5.3.3. Materiales

Para la asignatura de Materiales, los valores de acierto para el suspenso son máximos y estables a partir de un valor de C superior a 1000. Lo mismo ocurre para el acierto de los aprobados, que hasta el valor de 1000 (incluido), el grado de acierto para este grupo ronda valores entre el 100 y el 95 %.

De la *ilustración 12* se extrae que el valor máximo de aciertos en los suspensos se produce en un valor de C de 100000.

Por lo que respecta a la matriz de confusión de la *tabla 42*, se acierta 80 valores correspondientes al suspenso y 404 correspondientes al aprobado. Se predicen 141 datos como aprobados de forma incorrecta y 66 como suspensos también erróneamente.

Con los resultados que se obtienen de la matriz de confusión, se calculan los resultados que se muestran en la *tabla 43*.

El porcentaje de acierto que se obtiene para el grupo de suspensos es del 36%, y en cambio, para los aprobados es de un 86%. La precisión del modelo en los suspensos es también inferior al de los aprobados, 55% contra un 74 % de los aprobados. El acierto del total de datos se produce en un 71% de estos.

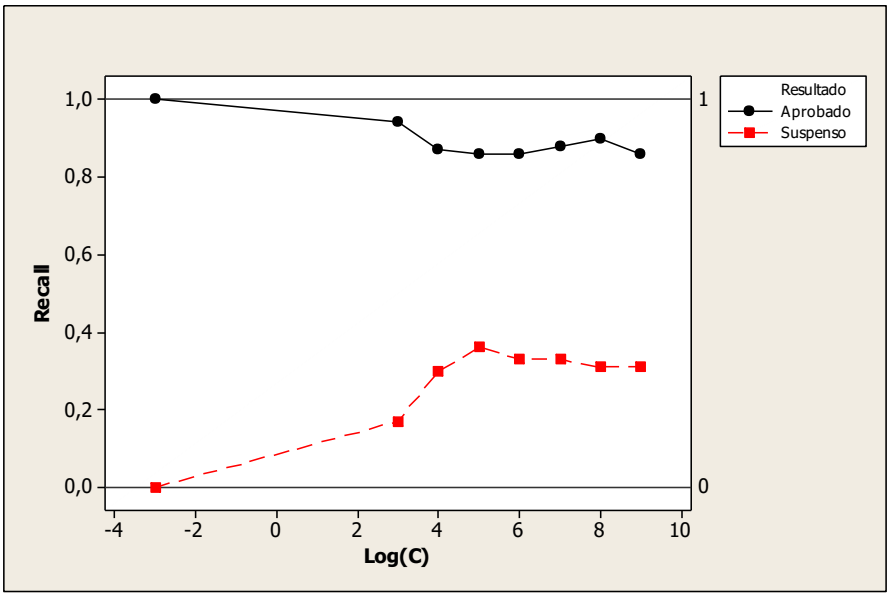


*Ilustración 12. Gráfica de Recalls en función del hiperparámetro C para Materiales.*

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | 80 | 141 |
| **Valor Real 1** | 66 | 404 |

*Tabla 42. Matriz de confusión de Recall máximo para Materiales.*

| C | Precisión Susp. | Precisión Aprob. | Recall Susp. | Recall Aprob. | Tasa Ac. |
|---|---|---|---|---|---|
| 100000 | 0.55 | 0.74 | 0.36 | 0.86 | 0.71 |

*Tabla 43. Tabla de resultados de Materiales. (SVM)*

### 5.3.4. EDOS

Graficando la evolución de los *Recalls* en función de C (*ilustración 13*) se obtiene que para la asignatura de EDOS el valor de los *Recalls* es constante para valores de C inferiores o iguales a 1000. En la que todo el acierto se produce en los aprobados de la asignatura y todo el error en los suspensos. A partir de un valor de 1000, se empieza a acertar en los suspensos y a cometer error en los aprobados.

Finalmente, se escoge el valor de C correspondiente al máximo de acierto para la clase suspenso. El valor del parámetro que obtiene el máximo es 10000, de forma similar a Electromagnetismo.

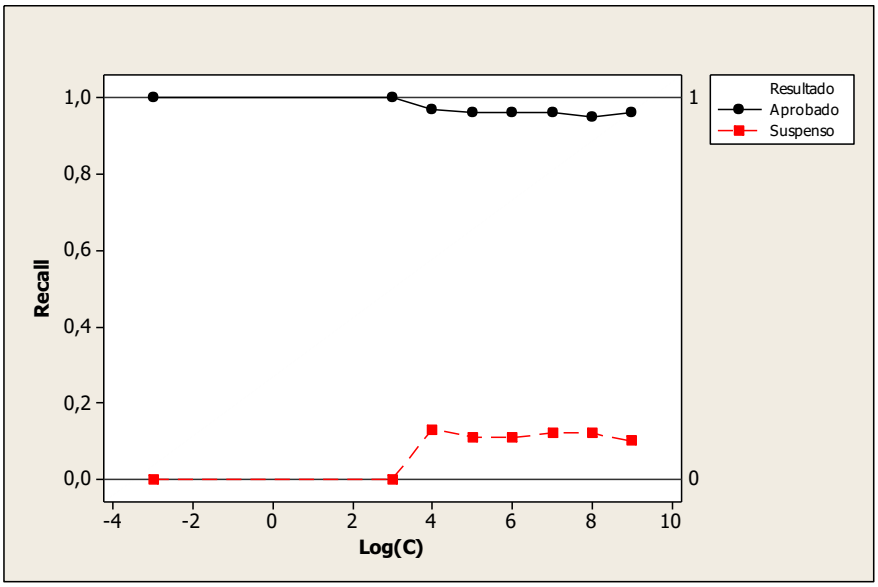


*Ilustración 13. Gráfica de Recalls en función del hiperparámetro C para EDOS.*

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | 18 | 124 |
| **Valor Real 1** | 16 | 527 |

*Tabla 44. Matriz de confusión de Recall máximo para EDOS.*

| C | Precisión Susp. | Precisión Aprob. | Recall Susp. | Recall Aprob. | Tasa Ac. |
|---|---|---|---|---|---|
| 10000 | 0.53 | 0.81 | 0.13 | 0.97 | 0.80 |

*Tabla 45. Tabla de resultados de EDOS. (SVM)*

De la matriz de confusión de la *tabla 44*, se observa que el modelo solamente acierta en 18 suspensos, para los aprobados acierta 527. Se asignan como suspensos de forma equivocada 16 datos. En cambio, se asignan de forma equivocada 124 datos como aprobado. Con el parámetro C anterior, se tabulan los resultados en la *tabla 45*. El grado de acierto máximo que se consigue es de un 13% para los suspensos y del el 97% de los aprobados de forma correcta. La precisión de los suspensos es de un 53% y de un 81 % para los aprobados. Los datos están muy desbalanceados, y la mayoría de los que se han utilizado para construir el hiperplano y sus márgenes pertenecen a la clase aprobado, por esta razón, la mayoría son clasificados como tal y el número de aciertos de esta clase es tan elevado respecto a la clase Suspenso.

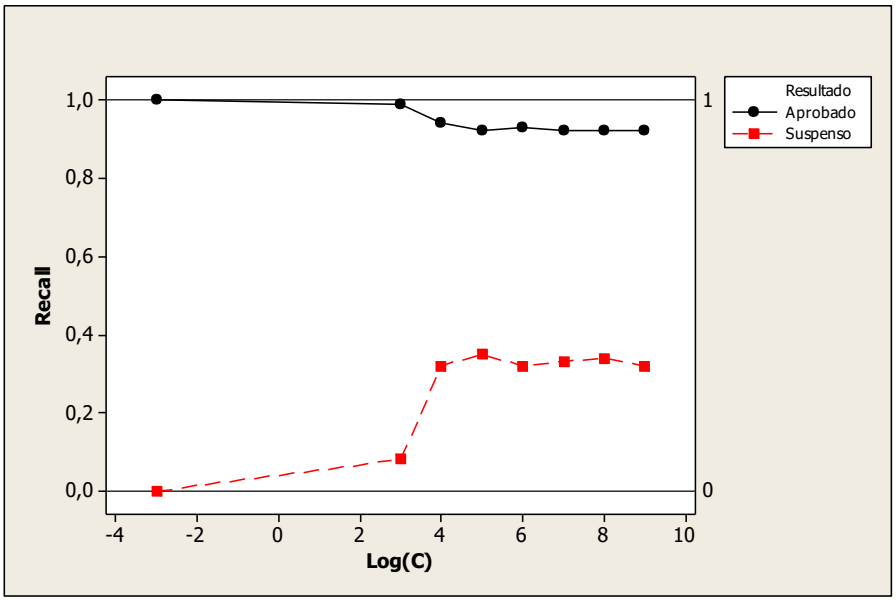


*Ilustración 14. Gráfica de Recalls en función del hiperparámetro C para Informática.*

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | 45 | 83 |
| **Valor Real 1** | 45 | 512 |

*Tabla 46. Matriz de confusión de Recall máximo para Informática.*

| C | Precisión Susp. | Precisión Aprob. | Recall Susp. | Recall Aprob. | Tasa Ac. |
|---|---|---|---|---|---|
| 10000 | 0.50 | 0.86 | 0.35 | 0.92 | 0.81 |

*Tabla 47. Tabla de resultados de Informática. (SVM)*

### 5.3.5. Informática

En la *ilustración 14* se muestra la gráfica de *Recalls* de los aprobados y de los suspensos para Informática en función de C, la asignatura presenta valores de *Recall* del Suspenso muy bajos si C presenta valores iguales o inferiores a 1000.

A partir de un valor de C de 10000, los aciertos de la clase aprobado son cercanos al 30 %. El máximo de estos valores aparece para un valor de C de 100000. Con este modelo se tabulan los resultados en la *tabla 47* y la matriz de confusión en la *tabla 46*.

El acierto máximo de la clase suspenso es de un 35% del total de los suspensos, para este modelo, el porcentaje de aprobados bien clasificados es del 92%. La precisión a la hora de asignar suspensos es de un 50 % exacto, es decir, el modelo asigna suspensos de forma

correcta e incorrecta por igual. Para los aprobados, la precisión es del 86%.

El número de aciertos en la clase suspenso es 45, y para los aprobados, de forma muy superior, es de 512, hay que tener en cuenta que el número de aprobados es también, muy superior respecto al de los suspensos. El modelo asigna de forma incorrecta 45 suspensos y 83 aprobados.

Hasta ahora, la probable no separación lineal y que los datos están desbalanceados, provocan que la fiabilidad se atribuya únicamente a la clase aprobado, por esta razón, el número de suspensos acertados y la precisión de esta clase corresponde a valores muy bajos.

### 5.3.6. Mecánica

Si se observa la evolución de los *Recalls* en la *ilustración 15*, se puede ver que el comportamiento de Mecánica es diferente al de las anteriores asignaturas ya que la curva referente a los suspensos toma valores iguales o superiores a la de los aprobados en todo momento.

Los valores de la curva de suspenso se mantienen constantes hasta un valor de 100 para el hiperparámetro C. Entonces, para saber qué modelo es el más adecuado, se puede ver la muy ligera pendiente negativa que se produce en la curva de aprobados, esto permite escoger como valor más acertado para C, 0.001.

Para el valor de C de 0.001 se muestra la matriz de confusión y los resultados en las *tablas 48 y 49*. El porcentaje de acierto máximo para los suspensos es del 82 %, este modelo obtiene un porcentaje de acierto para los aprobados de un 59%. La precisión al asignar suspensos es del 63% y al asignar aprobados es del 76%.

El número de suspensos correctamente asignados es de 256, y de aprobados correctamente asignados es de 214. Los datos que se asignan como suspenso pero que son aprobados en realidad son 148, y en cambio, los datos que son suspensos, pero que se asignan como aprobados, son 67.

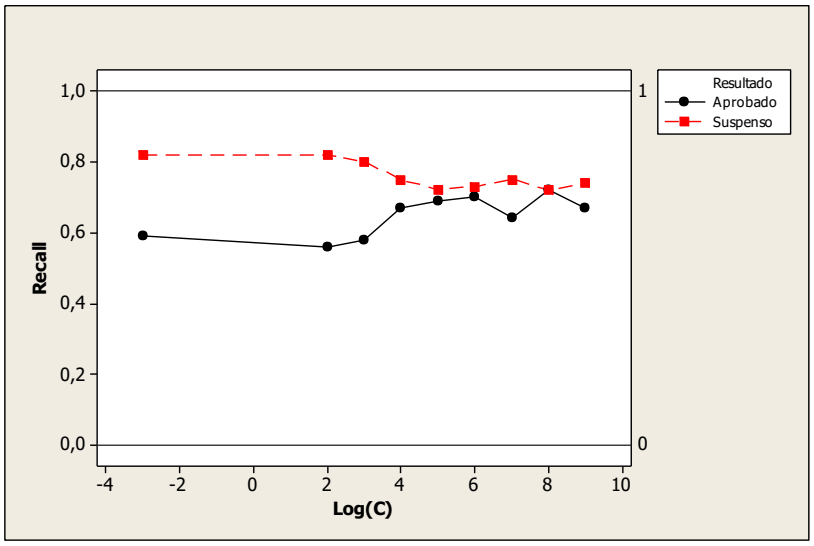


*Ilustración 15. Gráfica de Recalls en función del hiperparámetro C para Mecánica.*

| | **Valor predicho 0** | **Valor predicho 1** |
|---|---|---|
| **Valor Real 0** | 256 | 67 |
| **Valor Real 1** | 148 | 214 |

*Tabla 48. Matriz de confusión de Recall máximo para Mecánica.*

| C | Precisión Susp. | Precisión Aprob. | Recall Susp. | Recall Aprob. | Tasa Ac. |
|---|---|---|---|---|---|
| 0.001 | 0.63 | 0.76 | 0.82 | 0.59 | 0.69 |

*Tabla 49.Tabla de resultados de Mecánica. (SVM)*

### 5.3.7. Resumen SVM

En la *tabla 50* se muestra un resumen de los resultados para cada uno de los modelos empleados.

Para las asignaturas desbalanceadas, el número de aciertos de la clase suspenso es inferior al 50%, los porcentajes de acierto más bajos que se obtienen son los de las asignaturas de Métodos y Materiales, que son valores de acierto inferiores al 15%. Para estas mismas asignaturas, los valores de los *Recalls* de aprobados son los más elevados, con valores por encima del 95%.

Para las asignaturas desbalanceadas, los valores de C son muy elevados para poder equilibrar el número de vectores de cada tipo utilizados en el cálculo de los hiperplanos. Ya que escoger un valor elevado permite reducir el número de aprobados que se utiliza en el cálculo, reduciendo en menor medida el número de suspensos que se utilizan.

| | Recall Suspenso | Recall Aprobado | Tasa Aciertos | C |
|---|---|---|---|---|
| **Electromagnetismo** | 0.49 | 0.85 | 0.73 | 10000 |
| **Métodos** | 0.12 | 0.99 | 0.87 | 100000000 |
| **Materiales** | 0.36 | 0.86 | 0.71 | 100000 |
| **EDOS** | 0.13 | 0.97 | 0.80 | 10000 |
| **Informática** | 0.35 | 0.92 | 0.81 | 10000 |
| **Mecánica** | 0.82 | 0.59 | 0.69 | 0.001 |

*Tabla 50. Tabla resumen de los resultados de las SVM.*

La asignatura de Mecánica requiere de valores bajos de C para obtener resultados más satisfactorios en la predicción de suspensos. De igual forma, es la única asignatura que obtiene resultados exitosos respecto a los objetivos planteados.

La modelización de estos datos presenta dificultades debido a la distribución que estos ofrecen ya que los resultados sugieren que no existe apenas separabilidad lineal entre las dos clases (aprobado y suspenso). El hecho de trazar un plano para separar ambos conjuntos es un método demasiado simple para la compleja distribución de las dos clases. Es por ello, que la no separabilidad lineal de los datos afectada por la existencia de un grupo excesivamente mayoritario provoca una clara diferencia en el acierto de ambos grupos.

## 5.4. Comparativa de resultados

### 5.4.1. Electromagnetismo

En la *tabla 51* se muestran los resultados sobre el conjunto de *testing* de cada uno de los modelos para la asignatura de Electromagnetismo.

El modelo que recibe los resultados más satisfactorios para los objetivos planteados es el Árbol de Decisión. El modelo es capaz de predecir de forma acertada el 60% de los suspensos, un 10% más de acierto que el segundo mejor modelo (Regresión Logística). El incremento de acierto

| | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|
| **Regresión Logística** | 0.50 | 0.84 | 0.73 |
| **Árbol de Decisión** | 0.60 | 0.71 | 0.67 |
| **SVM** | 0.49 | 0.85 | 0.73 |

*Tabla 51. Tabla resumen de los resultados de los métodos predictivos empleados. (Electromagnetismo)*

del Árbol de Decisión en los suspensos penaliza el acierto de los aprobados, para la regresión logística existe un grado de acierto del 50% en la clase suspenso del conjunto de *testing* y de un 84% para los aprobados, en cambio, el aumento en el porcentaje de éxitos que produce el Árbol de Decisión afecta con un decremento de un 13% de acierto a la clasificación de la clase aprobado. Debido al desbalance entre el número existente de aprobados y suspensos, el acierto del modelo más satisfactorio es menor al de los otros dos modelos con un 67% de acierto global respecto al 73% de éxito global que se produce en el modelo de SVM y de Regresión Logística.

El resultado que se obtiene para los modelos de SVM y de Regresión Logística es prácticamente el mismo. Ambos modelos predicen con éxito casi el mismo número de suspensos y aprobados. Si se busca identificar qué modelo resulta más adecuado para cumplir con los objetivos, se escogería el modelo de Regresión Logística por tener un acierto mayor para los suspensos, sin embargo, el aprobado es predicho mejor en el modelo de SVM, y por ello, en el total de los resultados, el acierto también es superior (en la tabla aparecen los resultados redondeados pero el acierto más elevado es el de la SVM). Que el resultado sea tan similar para ambos grupos quiere decir que, en este caso, cambiar el método de separar los datos de forma lineal no influye, ya que por un método u otro los resultados son muy similares. Es posible que el conjunto de *testing* determine cuál de los dos modelos presenta mayor grado de éxito, es decir, dependiendo del conjunto que se utilice, el resultado puede favorecer a un modelo o a otro.

Indiscutiblemente, el modelo más satisfactorio para esta asignatura es el del Árbol de Decisión, independientemente del conjunto de *testing* en que se evalúe (siempre y cuando sea un conjunto de datos representativo). Que el Árbol de Decisión obtenga mejores resultados confirma que la separación de los datos no es lineal y que las subdivisiones que el Árbol lleva a cabo permiten detectar regiones más complejas, produciendo un acierto mayor para la clase Suspenso, que otros métodos no son capaces de alcanzar para la forma en que han sido empleados.

| | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|
| **Regresión Logística** | 0.09 | 0.99 | 0.87 |
| **Árbol de Decisión** | 0.33 | 0.89 | 0.82 |
| **SVM** | 0.12 | 0.99 | 0.87 |

*Tabla 52. Tabla resumen de los resultados de los métodos predictivos empleados. (Métodos Numéricos)*

### 5.4.2. Métodos Numéricos

Los resultados de los tres modelos sobre el conjunto de *testing* para Métodos Numéricos aparecen tabulados en la *tabla 52*.

La asignatura de Métodos Numéricos presenta un grado de desbalance todavía mayor al de la asignatura de Electromagnetismo. Por esta razón, clasificar los suspensos es, en este caso, todavía más complicado.

El grado de acierto máximo que se puede alcanzar para la clase suspenso es de un 33%. Este grado de acierto se da en el modelo del Árbol de Decisión, al igual que antes, este resulta ser el modelo más satisfactorio de los tres. El grado de éxito de las predicciones que se realizan para esta clase es muy inferior para los otros dos modelos. El modelo de Regresión Logística obtiene el grado de acierto más bajo, un 9% sobre el total de los suspensos del conjunto de *testing*. El modelo de SVM consigue clasificar de forma correcta el 12 % de los suspensos.

Los aprobados del conjunto de *testing* se clasifican de forma satisfactoria para todos los modelos, al igual que ha ocurrido con Electromagnetismo, conseguir un elevado grado de acierto en los suspensos provoca, para el Árbol de Decisión, un descenso en el acierto de los aprobados, aun así, el acierto en los aprobados sigue siendo elevado, un 89%. En los otros dos modelos, se catalogan de forma correcta el 99% de los aprobados, prácticamente todos. Al ser mayor el número de aprobados que de suspensos, incrementar en una unidad el porcentaje de éxitos en los aprobados respecto a los suspensos tiene una repercusión mucho mayor en el éxito del total de los datos, por ello el grado de éxito total es mayor en el modelo de SVM y de Regresión que en el del Árbol de Decisión.

Debido al bajo número de suspensos, puede parecer que el número de suspensos acertados por parte del árbol de decisión es muy superior al de los otros dos modelos, cuando en realidad, el

número de aciertos que tiene de más este modelo respecto al de SVM es de 19. Y a cambio, se dejan de clasificar de forma correcta 56 aprobados. A mayor desbalance, mayor penalización se da en el acierto del grupo mayoritario para una mejora en la predicción del grupo minoritario en el Árbol de Decisión.

Finalmente, para esta asignatura se escoge también el Árbol de Decisión como modelo idóneo y, en segundo lugar, el modelo de SVM por presentar la misma efectividad en los aprobados, pero superior a la de la Regresión Logística en los suspensos. Al igual que antes, las regiones que alcanza el Árbol de Decisión son mucho más complejas e inalcanzables por los otros dos métodos, y por ello se consigue detectar en mayor grado el patrón de los suspensos. Al igual que antes, los resultados confirman que los datos no presentan una separación lineal. El similar resultado que se obtiene por los métodos de separación lineal, indica que el método para llevar a cabo la separación no influye en los resultados.

### 5.4.3. Materiales

En la *tabla 53* se muestra un resumen de los resultados para cada uno de los modelos empleados en los datos de *testing* de la asignatura de Materiales.

El balance de los datos para Materiales es similar al de la asignatura de Electromagnetismo, pero los resultados que se consiguen para la clase suspenso presentan valores de acierto inferiores. Como ya se había comentado, esto puede ser debido a que los datos de la clase suspenso están más repartidos, es decir, existe una menor concentración de estos en torno a una región.

El grado de acierto más elevado para la clase suspenso se da en el modelo del Árbol de Decisión, el acierto máximo que se consigue es del 44% de los suspensos totales del conjunto de *testing*. Para el modelo de SVM, se consigue un acierto del 36% del total de los datos. El modelo con menor grado de acierto en los suspensos corresponde al de Regresión Logística, con un 31%.

El modelo de SVM, que consigue un acierto de un 5% de los suspensos superior respecto al modelo de Regresión, también consigue un porcentaje de acierto mayor al del modelo de Regresión en la clase aprobado. Esto no ocurre para el modelo del árbol de decisión, que tiene el grado de acierto más bajo de los tres a la hora de clasificar los aprobados, con un 75% de acierto.

El modelo que mejor actuación presenta en la totalidad de los datos de *testing* es el de SVM, que obtiene un 72% de éxito en la clasificación. Los otros dos modelos obtienen un acierto de un 68%.

| | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|
| **Regresión Logística** | 0.31 | 0.85 | 0.68 |
| **Árbol de Decisión** | 0.44 | 0.75 | 0.68 |
| **SVM** | 0.36 | 0.86 | 0.71 |

*Tabla 53. Tabla resumen de los resultados de los métodos predictivos empleados. (Materiales)*

Finalmente, el modelo que se adecúa mejor a la clasificación de la clase suspenso, y por tanto el que se escoge como mejor modelo, es el del Árbol de Decisión. Este método permite alcanzar regiones complejas que identifican una mayor cantidad de suspenso, los otros métodos emplean un mecanismo de división de los datos más simple y en este caso, eso no beneficia a los aciertos de los suspensos. El segundo mejor modelo es del SVM, aquí, el método de separación lineal que se utiliza si tiene influencia sobre los resultados, ya que este método clasifica mejor que el de Regresión.

Los resultados que se obtienen para esta asignatura no son del todo satisfactorios, ya que ningún modelo es capaz de predecir al menos el 50% de los suspensos presentes, y por tanto, la utilidad de cada uno de los modelos, a la hora de satisfacer los objetivos iniciales, es baja.

### 5.4.4. EDOS

EDOS es una asignatura con un elevado grado de desbalance entre en los aprobados y los suspensos. Por ello se espera que la predicción de los suspensos obtenga un bajo porcentaje de acierto. Por el contrario, se espera un buen grado de acierto a la hora de clasificar los aprobados del *testing*.

En la *tabla 54* se muestra el resultado de cada uno de los modelos. Si se mira el *Recall* de los suspensos, el modelo que presenta una mejor actuación es el del Árbol de Decisión, con un acierto del 35%. El grado de acierto de los otros dos modelos es muy inferior, con un valor de acierto del 13% para el modelo de SVM y de un 6% para el modelo de Regresión.

El grado de acierto del modelo de SVM para la clase suspenso es de más del doble del de la Regresión Logística. Y en cambio, al contrario de lo que se espera, el porcentaje de acierto que ambos presentan para la clase suspenso es el mismo, con un acierto de un 97%. Por lo que

| | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|
| **Regresión Logística** | 0.06 | 0.97 | 0.78 |
| **Árbol de Decisión** | 0.35 | 0.84 | 0.74 |
| **SVM** | 0.13 | 0.97 | 0.80 |

*Tabla 54. Tabla resumen de los resultados de los métodos predictivos empleados. (EDOS)*

respecta al modelo del Árbol de Decisión, la clasificación de los aprobados recibe la peor actuación de las tres, con un acierto del 84%.

Como es de esperar, el modelo con menor acierto en la clase mayoritaria presenta también el menor acierto en el total de los datos, este es el caso del Árbol de Decisión, que consigue acertar en la clasificación del 74% de los datos. Para los modelos restantes, la predicción de la clase mayoritaria es la misma, y el desempate lo genera la predicción en de la clase minoritaria, en este caso los suspensos, el modelo que consigue la mejor predicción es del SVM con un acierto del 80% de los datos. El modelo de Regresión acierta en la catalogación del 78% de los datos.

El modelo que mejor se adecua a los objetivos que se quieren cumplir es del Árbol de Decisión.. Si bien el resultado que se obtiene para los suspensos no es muy satisfactorio, el tipo de divisiones que el Árbol crea beneficia la clasificación de los Suspensos, y por ello el resultado es claramente superior al resultado de los otros dos modelos, que consiguen unas actuaciones muy pobres, destacando la baja fiabilidad del modelo de Regresión Logística.

Los resultados que se consiguen son los que se esperan, ya que la distribución de los datos de esta asignatura es parecida a la de Métodos Numéricos y por ello antes de visualizar los resultados de tiene una ligera idea de los que se puede obtener. Si se comparan los resultados que se obtienen para ambos modelos, se puede ver que el acierto de cada método es similar para las dos asignaturas.

### 5.4.5. Informática

Inicialmente se había comentado que el balance entre aprobados y suspensos para la asignatura de Informática es similar al de la asignatura de Métodos Numéricos y de EDOS. Pero los resultados que se obtienen con cada uno de los métodos para esta asignatura difieren al de las otras dos asignaturas.

| | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|
| **Regresión Logística** | 0.27 | 0.93 | 0.81 |
| **Árbol de Decisión** | 0.43 | 0.78 | 0.72 |
| **SVM** | 0.35 | 0.92 | 0.81 |

*Tabla 55. Tabla resumen de los resultados de los métodos predictivos empleados. (Informática)*

En la *tabla 55* aparecen tabulados los resultados de los tres modelos que se obtienen con cada uno de los métodos empleados.

El grado de acierto que presenta la asignatura en la clase suspenso es muy superior para cada uno de los métodos respecto a las asignaturas de Métodos Numéricos y EDOS. El acierto máximo para esta clase se consigue con el modelo del Árbol de Decisión, que obtiene un acierto del 43% del total del conjunto de *testing*. El acierto que se obtiene para los otros dos modelos es inferior y con valores de un 27% para el de Regresión y de un 35% para el modelo de SVM.

El grado de acierto para la clase aprobado es mayor en el modelo de Regresión Logística que en los otros dos, con un acierto del 93%. El modelo de SVM no dista del porcentaje de acierto anterior, el grado de éxito que éste consigue es del 92%. Finalmente, el modelo que consigue acertar en menor grado en la clasificación de los aprobados es el modelo del Árbol de Decisión con un 78% de acierto.

El global de datos se clasifica de forma muy similar para el modelo de SVM y de Regresión, ya que ambos modelos catalogan de forma correcta el 81% de los datos de *testing.* El modelo que clasifica peor el global, es el del Árbol de Decisión, con un acierto total del 72% de los datos, un 9% más de error que los otros dos modelos.

El modelo que consigue satisfacer los objetivos en mayor grado es el del Árbol de Decisión, aunque los valores de acierto de la clase suspenso sean inferiores al 50%, los resultados que se consiguen son mejor de los que en un principio cabía esperar si se observan los que se obtienen para las asignaturas de Métodos Numéricos y EDOS, dos asignaturas con un desbalance similar. En este caso realizando divisiones lineales ya se puede detectar al menos un cuarto de los suspensos, pero aún así, detectar regiones más complejas a través de las subdivisiones que el Árbol realiza ayuda a identificar un mayor número de suspensos.

| | Recall Suspenso | Recall Aprobado | Tasa Aciertos |
|---|---|---|---|
| **Regresión Logística** | 0.76 | 0.60 | 0.68 |
| **Árbol de Decisión** | 0.84 | 0.46 | 0.64 |
| **SVM** | 0.82 | 0.59 | 0.69 |

*Tabla 56. Tabla resumen de los resultados de los métodos predictivos empleados. (Mecánica)*

### 5.4.6. Mecánica

Los resultados que se consiguen para la asignatura de Mecánica son satisfactorios para todos los modelos que se emplean. Todos los modelos son capaces de acertar en la clasificación de los suspensos en más del 75% de estos en el conjunto de *testing*.

En la *tabla 56* se enseñan los resultados que cada uno de los modelos consigue.

Los resultados muestran, al igual que en el resto de las asignaturas, que el modelo de predicción que mayor acierto obtiene a la hora de clasificar los suspensos es el modelo referente al Árbol de Decisión. En este caso, el modelo consigue un acierto del 84% del total de datos de la clase Suspenso. Los resultados que se obtienen empleando los otros métodos son también similares, el modelo de SVM obtiene un grado de acierto del 82% y el de Regresión de un 76% para este grupo.

Por lo que respecta a la predicción de la clase Aprobado, el modelo más acertado es el de la Regresión Logística, con un acierto del 60%. Seguidamente, el modelo con segundo mayor grado de acierto de aprobados es el de SVM, con un 59%. Y finalmente, el que consigue los peores resultados en la predicción de esta clase, es el modelo del Árbol de Decisión, con un 46%.

En esta asignatura los datos están balanceados y, por tanto, el acierto más elevado lo presenta el modelo que obtiene una mejor combinación de acierto entre los aprobados y suspensos. El modelo con la mejor combinación es el de SVM, que consigue un acierto global de un 69%. El modelo de Regresión obtiene un acierto muy similar, un 68%. El que presenta un grado de éxito global más bajo es el modelo del Árbol de Decisión, con un 64% de éxito.

En este caso se tienen dos modelos que predicen de forma muy similar los suspensos del conjunto de *testing,* el modelo del Árbol de Decisión y el de SVM. Por esta razón otro factor a

tener en cuenta en la elección del modelo será la precisión. Para ello, si se compara los *Recalls* del aprobado de estas dos clases, se puede ver que se comete un elevado, y mayor, porcentaje de error en el modelo del Árbol de Decisión que en el de SVM. En este caso, a diferencia de los anteriores, no es el que presenta mayor acierto en la clase suspenso, el modelo que satisface mejor los objetivos, sino que es el de SVM.

Los resultados que se obtienen para todos los modelos son muy similares entre sí, esto se debe a que, en este caso, como el Árbol de Decisión tiene una profundidad de 1, el hecho de obtener un árbol así sugiere que la división lineal funciona mejor para esta asignatura y también que, por lo que se ve en los resultados, la partición hecha por la regresión y la SVM es mejor. Es decir, que para datos linealmente separables el árbol con un nodo es peor predictor.

### 5.4.7. Resumen de los modelos escogidos

En la *tabla 57* se enseña la tabla resumen con los resultados definitivos de cada uno de los modelos escogidos.

Los modelos que permiten obtener un mejor balance de resultados para la clase aprobado y suspenso son los que se crean empleando Árboles de Decisión para las asignaturas desbalanceadas, para Mecánica, en cambio, el mejor balance se obtiene con el modelo de SVM. Con los resultados que se consiguen, se puede ver que para aquellos datos que presentan un elevado grado de desbalance, el Árbol de Decisión es un método eficaz de predicción, ya que gracias a la profundidad que se le otorga al Árbol, se pueden explorar múltiples regiones más complejas sobre el conjunto de los datos que otorgan un grado de precisión que los modelos de Regresión y SVM empleados no pueden alcanzar.

Por lo que respecta a los resultados que se obtienen con el modelo de SVM, se deduce que los datos referentes a la asignatura de Mecánica presentan dos regiones diferenciadas donde en cada una de ellas existe una concentración de datos de una clase que predomina sobre la otra. Ya que con situar un hiperplano entre ambas regiones basta para obtener resultados satisfactorios en la predicción de los aprobados y los suspensos del conjunto de *testing*.

Los únicos modelos que consiguen superar el 50% de acierto de los suspensos son Electromagnetismo y Mecánica, se puede afirmar que estos resultados son satisfactorios para esta clase. El resto de las asignaturas consigue un porcentaje de acierto bajo en esta clase, con unos valores entre el 30 y el 45% de acierto. Los valores más elevados en la clase aprobado se

| Asignatura | Modelo | Recall Suspenso | Recall Aprobado | Tasa Aciertos. |
|---|---|---|---|---|
| **Electromagnetismo** | Árbol | 0.60 | 0.71 | 0.67 |
| **Métodos** | Árbol | 0.33 | 0.89 | 0.82 |
| **Materiales** | Árbol | 0.44 | 0.75 | 0.68 |
| **EDOS** | Árbol | 0.35 | 0.84 | 0.74 |
| **Informática** | Árbol | 0.43 | 0.78 | 0.72 |
| **Mecánica** | SVM | 0.82 | 0.59 | 0.69 |

*Tabla 57. Tabla resumen de los resultados para cada uno de los modelos escogidos.*

obtienen para los modelos que consiguen las peores actuaciones sobre los suspensos, aun así, todos los resultados que se obtienen en la clasificación de esta clase son satisfactorios. Para las asignaturas que presentan un mayor grado de desbalance (Métodos Numéricos, EDOS e Informática), se obtiene un porcentaje de éxito global superior al 70%, ya que la obtención de una gran cantidad de acierto en los aprobados para las asignaturas desbalanceadas garantiza buenos resultados en el cómputo global de los datos.

Finalmente es interesante valorar si las SVM que se han utilizado en este proyecto respecto a la Regresión empleada, merecen la pena desde el punto de vista del coste computacional. Ya que ambos métodos buscan la separación lineal de los datos, pero el método de hacerlo es diferente para cada uno y la complejidad del SVM lo convierte en un método más costoso en ese aspecto. Por lo visto, con los resultados obtenidos, el SVM no es capaz de incrementar de forma destacada la tasa de acierto en algunas asignaturas y por ello, se concluye que su uso, en este caso, no merece la pena.

# 6. Presupuesto

Los costes referentes a este trabajo pueden desglosarse en costes de personal y en los costes del equipamiento utilizado. La suma de los costes totales corresponde al valor de cada una de las horas empleadas por el personal más el valor de los equipos de trabajo.

## 6.1. Coste de personal

Los costes de personal involucran las horas invertidas por el analista para el desarrollo del proyecto y también, las horas que se dedican para guiar y para resolver problemas por parte del director del proyecto.

Las horas que se dedican al proyecto suponen un coste variado. El coste de las horas varía en función de la etapa en la que se inviertan. Las etapas que se consideran son: formación e investigación, análisis y presentación. Las horas que el director del proyecto dedica se cuantifican aparte de las etapas nombradas.

La primera etapa hace referencia a la etapa de adquisición de conocimientos de la problemática a tratar, de la selección de la metodología a emplear y de la obtención de los conocimientos necesarios para la comprensión e implementación de los diferentes métodos de predicción.

La etapa de análisis es la etapa de implementación del modelo, asimilación del mismo y de comprensión de cada uno de los resultados que se consiguen.

Por último, la etapa final del trabajo es la de presentación, en esta etapa, se dedica el tiempo a transformar los resultados a un formato sencillo para facilitar su comprensión y a realizar la redacción de la memoria.

## 6.2. Costes de equipamiento

El único coste que aparece en este apartado, es el coste asociado al ordenador que se emplea para llevar a cabo todas las tareas del proyecto. Las herramientas que se utilizan son de uso libre, de manera que suponen ningún coste adicional.

| COSTES DE PERSONAL | | | |
|---|---|---|---|
| **Etapa** | **Precio/hora** | **Horas Dedicadas** | **Coste Total** |
| Formación e Investigación | 40€ | 120 | 4800€ |
| Análisis | 50€ | 100 | 5000€ |
| Presentación | 30€ | 60 | 1800€ |
| Ayudas del director | 100€ | 12 | 1200€ |
| **COSTES DE EQUIPAMIENTO** | | | 249.33€ |
| **COSTE FINAL DEL PROYECTO** | | | 13049.33€ |

*Tabla 58. Tabla de costes del proyecto.*

El valor inicial del ordenador era de 1600€, considerando un coste de mantenimiento de un 10% anual respecto a su precio inicial por un uso de 1200 horas al año, y considerando que el ordenador se utiliza en el 90% del tiempo que se dedica al proyecto.

$$300h \times 0.9 \times \frac{1600€ \times 0.1}{1200h} = 36€$$

Al coste anterior se añade también la amortización del ordenador, considerando un uso del ordenador de 300 días al año, durante 3 años respecto al momento de su adquisición, y un uso del mismo durante el proyecto de 120 días.

$$120d \times \frac{1600€}{3años \times 300d} = 213.33€$$

El coste total que deriva del ordenador es la suma de los dos calculados anteriormente.

$$36€ + 213.33€ = 249.33€$$

En la tabla X se muestran los costes que se asignan a cada elemento, el coste final que se estima para este proyecto es de 13049.33€.

# 7. Impacto Medioambiental

Este proyecto no provoca apenas un impacto medioambiental sobre el entorno, todas las herramientas que posibilitan el proyecto se emplean a través de un ordenador. La realización de este proyecto emplea una cantidad mínima de papel insuficiente como para ser considerada en esta etapa.

Desde el punto de vista energético, se tiene en cuenta el uso de redes a través de *routers* y la utilización de fuentes eléctricas de alimentación para garantizar la operatividad de los recursos que se emplean. También se consideran las fuentes de energía lumínica del entorno de trabajo, aun así, estas fuentes se emplean únicamente durante el periodo nocturno.

# 8. Planificación

En este apartado se muestra la planificación que se ha llevado a cabo a lo largo del proyecto hasta su finalización, en la *tabla 59* se muestra el diagrama de *Gantt* que integra cada una de las etapas desde el inicio hasta la presentación del proyecto.

| | | Año 2019 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Etapas** | **Actividades** | **Feb** | | **Mar** | | **Abr** | | **May** | | **Jun** | | **Jul** | |
| **Planteamiento del proyecto** | Definición de objetivos | | | | | | | | | | | | |
| | Establecimiento de una metodología | | | | | | | | | | | | |
| **Manipulación de datos** | Familiarización con la librería Pandas | | | | | | | | | | | | |
| | Tratamiento y limpieza de datos | | | | | | | | | | | | |
| | Obtención de conocimiento previo | | | | | | | | | | | | |
| **Creación y validación de modelos** | Estudio de algoritmos | | | | | | | | | | | | |
| | Familiarización con Scikit-Learn | | | | | | | | | | | | |
| | Aplicación de algoritmos | | | | | | | | | | | | |
| | Análisis de resultados | | | | | | | | | | | | |
| **Memoria** | Redacción de la memoria | | | | | | | | | | | | |
| | Presupuesto e impacto medioambiental | | | | | | | | | | | | |
| | Conclusiones | | | | | | | | | | | | |
| **Presentación** | Presentación del proyecto | | | | | | | | | | | | |

*Tabla 59. Diagrama de Gantt del proyecto.*

# 9. Conclusiones

Una vez se completa el proyecto, se considera que cada uno de los objetivos se satisface en mayor o menor grado. Aun así, en el cómputo global, la sensación presente es satisfactoria por los resultados y por el conocimiento adquirido, que resulta necesario para poder completar cada una de las etapas.

Se consigue aplicar la metodología *CRISP-DM,* dentro de la cuál, el conocimiento previo que se adquiere es satisfactorio al conseguir tener una idea de la distribución de los datos y del grado de dificultad que cada asignatura supone al alumno a la hora de cursarla.

El resultado del estudio y de la comparación de cada uno de los métodos no es definitivo ya que existen múltiples formas de emplear cada uno de los métodos que consiguen obtener resultados mejores, por ello se cree que en este apartado no se puede afirmar con seguridad que un método sea más efectivo que otro a la hora de llevar a cabo una predicción sobre el conjunto de datos disponible, en este caso resulta que el método que sale más perjudicado es la Regresión Logística que seguramente, empleado de otra forma los resultados podrían ser muy diferentes.

A pesar de no ser definitivos, los resultados que se obtienen son satisfactorios porque estos sugieren que lo más probable es que para muchas asignaturas del *Q3* los datos no sean linealmente separables y eso ayuda a escoger qué modelos testear en futuros trabajos. El hecho de que la tasa de aciertos no sea alta sugiere que a lo mejor los datos no contienen más información, aunque habría que probar otros métodos.

Gracias a la terminología expuesta con anterioridad se consigue entender cada uno de los resultados que los modelos ofrecen y, por tanto, la sensación en este aspecto es también, satisfactoria.

Finalmente, el objetivo secundario de conseguir un grado de acierto correcto a la hora de predecir los suspensos solo se consigue completar para los modelos de las asignaturas de Electromagnetismo y de Mecánica que ofrecen un grado de acierto superior al 50%. El grado de desbalance y la distribución de los datos que presentan las asignaturas restantes dificulta la obtención de resultados satisfactorios para la clase suspenso.

## Conclusión Personal

Como conclusión personal de este trabajo, se quiere destacar el descubrimiento de las múltiples

posibilidades que el mundo de la Minería de Datos y el *Machine Learning* ofrece, así como la utilidad de los métodos predictivos empleados para facilitar la toma de decisiones en este, pero también en otros ámbitos. Además, es necesario recalcar las habilidades que la implementación de cada línea de código supone y destacar también la agilidad que se adquiere con Python gracias a la utilización reiterada de los diferentes métodos para poder completar cada uno de los apartados que integran el proyecto.

## Trabajos futuros

El final de este proyecto no es cerrado ya que todo lo que se realiza hasta el momento tiene una continuidad. Gracias a los fragmentos de código del anexo, los análisis se pueden replicar de forma sencilla para realizar una profundización a un nivel superior respecto al actual. Cada uno de los métodos que se emplea puede ser mejorado para obtener unos resultados más fiables y precisos.

En el caso de la Regresión Logística, se han empleado variables que pueden no ser significativas en el modelo, y por ello los resultados que se consiguen en la predicción del suspenso son tan bajos. También, al tener variables discretas que presentan un alto grado de correlación con otras, una posible mejora en los resultados es la incorporación de nuevas variables que son el producto de las discretas con las continuas correlacionadas. Las variables también pueden requerir de transformaciones para alcanzar mejoras, las más típicas son la aplicación de logaritmos, exponenciales y cuadrados entre otras.

Los modelos de SVM que se emplean en este proyecto realizan la separación de los datos a través de hiperplanos lineales. Pero los SVM pueden implementar también, hiperplanos separadores no lineales a través de lo que se conoce como función *Kernel*. La función *Kernel* aplica una transformación sobre los datos no separables linealmente a un espacio en la que los datos si son separables de forma lineal. Es posible que la aplicación de la función *Kernel* sea capaz de mejorar los resultados de éxito bajos que se consiguen para los suspensos en algunas asignaturas.

Este proyecto puede ser prolongado también al empleo de otras técnicas de predicción muy utilizadas. El siguiente método predictivo a emplear podrían ser las Redes Neuronales, un método predictivo que es capaz de construir modelos no lineales

# Bibliografía

[1] Olson, David L. et al. *Advanced Data Mining Techniques*. Berlin: Springer, 2008. 169 p. ISBN 978-3-54-076916-3

[2] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, Inc, USA (24 de marzo de 2017). 566 p. ISBN: 978-1491962299

[3] Varios autores. *KDNuggets*. Dirección Web: https://www.kdnuggets.com/

[4] *Scikit-learn: Machine learning in Python* (documentación *Scikit-Learn*). Dirección Web: https://scikit-learn.org/stable/index.html

[5] *Python Data Analysis Library* (documentación de la librería *Pandas*. Dirección Web: https://pandas.pydata.org/

[6] Curso Completo de *Machine Learning*: *Data Science*. Dirección Web: https://www.udemy.com/

# Anexo

## Suppor Vector Machine

| C | Recall 0 | Support 0 | Recall 1 | Support 1 | Vectores de soporte |
|---|---|---|---|---|---|
| 0,001 | 0,16 | 221 | 0,97 | 464 | 1063 |
| 0,01 | 0,41 | 221 | 0,87 | 464 | 995 |
| 0,1 | 0,44 | 221 | 0,88 | 464 | 963 |
| 1 | | 221 | | 464 | |
| 10 | 0,44 | 221 | 0,87 | 464 | 959 |
| 100 | | 221 | | 464 | |
| 1000 | 0,44 | 221 | 0,85 | 464 | 937 |
| 10000 | 0,49 | 221 | 0,85 | 464 | 742 |
| 100000 | 0,47 | 221 | 0,85 | 464 | 648 |
| 1000000 | 0,46 | 221 | 0,84 | 464 | 646 |
| 10000000 | 0,46 | 221 | 0,86 | 464 | 637 |
| 10000000 | 0,46 | 221 | 0,84 | 464 | 633 |
| 100000000 | 0,47 | 221 | 0,84 | 464 | 634 |

*Tabla 60. Tabla de resultados en función del hiperparámetro C. (SVM: Electromagnetismo)*

| C | Recall 0 | Support 0 | Recall 1 | Support 1 | Vectores de soporte |
|---|---|---|---|---|---|
| 0,001 | 0 | 91 | 1 | 594 | 419 |
| 0,01 | | 91 | | 594 | |
| 0,1 | | 91 | | 594 | |
| 1 | | 91 | | 594 | |
| 10 | | 91 | | 594 | |
| 100 | | 91 | | 594 | |
| 1000 | 0 | 91 | 1 | 594 | 409 |
| 10000 | 0,05 | 91 | 0,99 | 594 | 395 |
| 100000 | 0,04 | 91 | 1 | 594 | 369 |
| 1000000 | 0,09 | 91 | 0,99 | 594 | 348 |
| 10000000 | 0,10 | 91 | 0,99 | 594 | 356 |
| 100000000 | 0,12 | 91 | 0,99 | 594 | 357 |
| 1000000000 | 0,06 | 91 | 0,99 | 594 | 360 |

*Tabla 61. Tabla de resultados en función del hiperparámetro C. (SVM: Métodos Numéricos)*

| C | Recall 0 | Support 0 | Recall 1 | Support 1 | Vectores de soporte |
|---|---|---|---|---|---|
| 0,001 | 0 | 215 | 1 | 470 | 996 |
| 0,01 | | 215 | | 470 | |
| 0,1 | | 215 | | 470 | |
| 1 | | 215 | | 470 | |
| 10 | | 215 | | 470 | |
| 100 | | 215 | | 470 | |
| 1000 | 0,17 | 215 | 0,94 | 470 | 959 |
| 10000 | 0,3 | 215 | 0,87 | 470 | 786 |
| 100000 | 0,36 | 215 | 0,86 | 470 | 711 |
| 1000000 | 0,33 | 215 | 0,86 | 470 | 703 |
| 10000000 | 0,33 | 215 | 0,88 | 470 | 703 |
| 10000000 | 0,31 | 215 | 0,9 | 470 | 698 |
| 100000000 | 0,31 | 215 | 0,86 | 470 | 699 |

*Tabla 62. Tabla de resultados en función del hiperparámetro C. (SVM: Materiales)*

| C | Recall 0 | Support 0 | Recall 1 | Support 1 | Vectores de soporte |
|---|---|---|---|---|---|
| 0,001 | 0 | 142 | 1 | 543 | 636 |
| 0,01 | | 142 | | 543 | |
| 0,1 | | 142 | | 543 | |
| 1 | | 142 | | 543 | |
| 10 | | 142 | | 543 | |
| 100 | | 142 | | 543 | |
| 1000 | 0 | 142 | 1 | 543 | 631 |
| 10000 | 0,13 | 142 | 0,97 | 543 | 569 |
| 100000 | 0,11 | 142 | 0,96 | 543 | 505 |
| 1000000 | 0,11 | 142 | 0,96 | 543 | 496 |
| 10000000 | 0,12 | 142 | 0,96 | 543 | 501 |
| 10000000 | 0,12 | 142 | 0,95 | 543 | 511 |
| 100000000 | 0,1 | 142 | 0,96 | 543 | 514 |

*Tabla 63. Tabla de resultados en función del hiperparámetro C. (SVM: EDOS)*

| C | Recall 0 | Support 0 | Recall 1 | Support 1 | Vectores de soporte |
|---|---|---|---|---|---|
| 0,001 | 0 | 128 | 1 | 557 | 702 |
| 0,01 | | 128 | | 557 | |
| 0,1 | | 128 | | 557 | |
| 1 | | 128 | | 557 | |
| 10 | | 128 | | 557 | |
| 100 | | 128 | | 557 | |
| 1000 | 0,08 | 128 | 0,99 | 557 | 687 |
| 10000 | 0,32 | 128 | 0,94 | 557 | 582 |
| 100000 | 0,35 | 128 | 0,92 | 557 | 540 |
| 1000000 | 0,32 | 128 | 0,93 | 557 | 523 |
| 10000000 | 0,33 | 128 | 0,92 | 557 | 515 |
| 100000000 | 0,34 | 128 | 0,92 | 557 | 515 |
| 1000000000 | 0,32 | 128 | 0,92 | 557 | 510 |

*Tabla 64. Tabla de resultados en función del hiperparámetro C. (SVM: Informática)*

| C | Recall 0 | Support 0 | Recall 1 | Support 1 | Vectores de soporte |
|---|---|---|---|---|---|
| 0,001 | 0,82 | 323 | 0,59 | 362 | 1241 |
| 0,01 | | 323 | | 362 | |
| 0,1 | | 323 | | 362 | |
| 1 | | 323 | | 362 | |
| 10 | | 323 | | 362 | |
| 100 | 0,82 | 323 | 0,56 | 362 | 1111 |
| 1000 | 0,8 | 323 | 0,58 | 362 | 1075 |
| 10000 | 0,75 | 323 | 0,67 | 362 | 839 |
| 100000 | 0,72 | 323 | 0,69 | 362 | 765 |
| 1000000 | 0,73 | 323 | 0,7 | 362 | 756 |
| 10000000 | 0,75 | 323 | 0,64 | 362 | 750 |
| 100000000 | 0,72 | 323 | 0,72 | 362 | 763 |
| 1000000000 | 0,74 | 323 | 0,67 | 362 | 769 |

*Tabla 65. Tabla de resultados en función del hiperparámetro C. (SVM: Mecánica)*

## Árboles de Decisión

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,67737226 | 0,67007299 | 0,67883212 | 0,66861314 | 0,66569343 | 0,68175182 | 0,64379562 | 0,65547445 | 0,65839416 | 0,64379562 | 0,64525547 | 0,64379562 | 0,64233577 |
| 2 | 0,67737226 | 0,67007299 | 0,67883212 | 0,66861314 | 0,66277372 | 0,68467153 | 0,64817518 | 0,64233577 | 0,6350365 | 0,65255474 | 0,64087591 | 0,6540146 | 0,6540146 |
| 3 | 0,67737226 | 0,67007299 | 0,67883212 | 0,66861314 | 0,66861314 | 0,6919708 | 0,66423358 | 0,65839416 | 0,6379562 | 0,64963504 | 0,64525547 | 0,65255474 | 0,66131387 |
| 5 | 0,67737226 | 0,67007299 | 0,67883212 | 0,66569343 | 0,66277372 | 0,66277372 | 0,65985401 | 0,65109489 | 0,65255474 | 0,65255474 | 0,65109489 | 0,65839416 | 0,65985401 |
| 7 | 0,67737226 | 0,67007299 | 0,67883212 | 0,67153285 | 0,67007299 | 0,65547445 | 0,6379562 | 0,65693431 | 0,66277372 | 0,65985401 | 0,65547445 | 0,66277372 | 0,66131387 |
| 10 | 0,67737226 | 0,67007299 | 0,67883212 | 0,67007299 | 0,68029197 | 0,66131387 | 0,64525547 | 0,64233577 | 0,64525547 | 0,64525547 | 0,64525547 | 0,64525547 | 0,64525547 |
| 15 | 0,67737226 | 0,67007299 | 0,67883212 | 0,67445255 | 0,6919708 | 0,65985401 | 0,6540146 | 0,64963504 | 0,6540146 | 0,6540146 | 0,64963504 | 0,6540146 | 0,64963504 |
| 20 | 0,67737226 | 0,67007299 | 0,67883212 | 0,70948905 | 0,71970803 | 0,70072993 | 0,69343066 | 0,69635036 | 0,69635036 | 0,69343066 | 0,69343066 | 0,69635036 | 0,69343066 |
| 25 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71824818 | 0,70218978 | 0,70364964 | 0,70364964 | 0,70364964 | 0,70364964 | 0,70364964 | 0,70364964 | 0,70364964 |
| 30 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71678832 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 |
| 35 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71678832 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 |
| 40 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 |
| 50 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 |

*Tabla 66.Scores para la asignatura de Electromagnetismo en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,86715328 | 0,86569343 | 0,86861314 | 0,86277372 | 0,84817518 | 0,83065693 | 0,76934307 | 0,77664234 | 0,77518248 | 0,77518248 | 0,77080292 | 0,77518248 | 0,77664234 |
| 2 | 0,86715328 | 0,86569343 | 0,86861314 | 0,86277372 | 0,84233577 | 0,81751825 | 0,76788321 | 0,76642336 | 0,75474453 | 0,7620438 | 0,75766423 | 0,77226277 | 0,76788321 |
| 3 | 0,86715328 | 0,86569343 | 0,86861314 | 0,86131387 | 0,84963504 | 0,82335766 | 0,7810219 | 0,78394161 | 0,76934307 | 0,79854015 | 0,78394161 | 0,79124088 | 0,78686131 |
| 5 | 0,86715328 | 0,86569343 | 0,86861314 | 0,86131387 | 0,84379562 | 0,8379562 | 0,80437956 | 0,80145985 | 0,80145985 | 0,80291971 | 0,80291971 | 0,8 | 0,8 |
| 7 | 0,86715328 | 0,86569343 | 0,86569343 | 0,86277372 | 0,8540146 | 0,84817518 | 0,82189781 | 0,80875912 | 0,82043796 | 0,82043796 | 0,82335766 | 0,82627737 | 0,82627737 |
| 10 | 0,86715328 | 0,87591241 | 0,86277372 | 0,86861314 | 0,85985401 | 0,84379562 | 0,82189781 | 0,82189781 | 0,82189781 | 0,8189781 | 0,8189781 | 0,82189781 | 0,82189781 |
| 15 | 0,86715328 | 0,87591241 | 0,86277372 | 0,85109489 | 0,85985401 | 0,84963504 | 0,84963504 | 0,84963504 | 0,84963504 | 0,84963504 | 0,84963504 | 0,84963504 | 0,84963504 |
| 20 | 0,86715328 | 0,87591241 | 0,87153285 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 |
| 25 | 0,86715328 | 0,87591241 | 0,87153285 | 0,85985401 | 0,85985401 | 0,85985401 | 0,85985401 | 0,85985401 | 0,85985401 | 0,85985401 | 0,85985401 | 0,85985401 | 0,85985401 |
| 30 | 0,86715328 | 0,87591241 | 0,87591241 | 0,86423358 | 0,86423358 | 0,86423358 | 0,86423358 | 0,86423358 | 0,86423358 | 0,86423358 | 0,86423358 | 0,86423358 | 0,86423358 |
| 35 | 0,86715328 | 0,87591241 | 0,87737226 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 | 0,86569343 |
| 40 | 0,86715328 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 |
| 45 | 0,86715328 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 |
| 50 | 0,86715328 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 | 0,87591241 |

*Tabla 67. Scores para la asignatura de Métodos Numéricos en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,68613139 | 0,67153285 | 0,67153285 | 0,67445255 | 0,6729927 | 0,63211679 | 0,64379562 | 0,62919708 | 0,63211679 | 0,62919708 | 0,62481752 | 0,62919708 | 0,61605839 |
| 2 | 0,68613139 | 0,67153285 | 0,67153285 | 0,67591241 | 0,68467153 | 0,63211679 | 0,62773723 | 0,6379562 | 0,62189781 | 0,6350365 | 0,62189781 | 0,63357664 | 0,63357664 |
| 3 | 0,68613139 | 0,67153285 | 0,67153285 | 0,67737226 | 0,68321168 | 0,6379562 | 0,64379562 | 0,64379562 | 0,64963504 | 0,6379562 | 0,64963504 | 0,6379562 | 0,64379562 |
| 5 | 0,68613139 | 0,67153285 | 0,67153285 | 0,67445255 | 0,66861314 | 0,63649635 | 0,63941606 | 0,64525547 | 0,64233577 | 0,64379562 | 0,6379562 | 0,64963504 | 0,64671533 |
| 7 | 0,68613139 | 0,67153285 | 0,67153285 | 0,67445255 | 0,6729927 | 0,63649635 | 0,6379562 | 0,63211679 | 0,62627737 | 0,64233577 | 0,63357664 | 0,64233577 | 0,6379562 |
| 10 | 0,68613139 | 0,67153285 | 0,67153285 | 0,67007299 | 0,68029197 | 0,62773723 | 0,62919708 | 0,62919708 | 0,62919708 | 0,62919708 | 0,62773723 | 0,62919708 | 0,63357664 |
| 15 | 0,68613139 | 0,67153285 | 0,67153285 | 0,67007299 | 0,65547445 | 0,64963504 | 0,65109489 | 0,64963504 | 0,65109489 | 0,64963504 | 0,64963504 | 0,64963504 | 0,64963504 |
| 20 | 0,68613139 | 0,67153285 | 0,67153285 | 0,66277372 | 0,65547445 | 0,65985401 | 0,65985401 | 0,65985401 | 0,65985401 | 0,65985401 | 0,65985401 | 0,65985401 | 0,65985401 |
| 25 | 0,68613139 | 0,67153285 | 0,67153285 | 0,68759124 | 0,67883212 | 0,67883212 | 0,67883212 | 0,67883212 | 0,67883212 | 0,67883212 | 0,67883212 | 0,67883212 | 0,67883212 |
| 30 | 0,68613139 | 0,67153285 | 0,67153285 | 0,69489051 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 |
| 35 | 0,68613139 | 0,67153285 | 0,67153285 | 0,68613139 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 |
| 40 | 0,68613139 | 0,67153285 | 0,67153285 | 0,68613139 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 | 0,68467153 |
| 45 | 0,68613139 | 0,67153285 | 0,67153285 | 0,68613139 | 0,68467153 | 0,69343066 | 0,69343066 | 0,69343066 | 0,69343066 | 0,69343066 | 0,69343066 | 0,69343066 | 0,69343066 |
| 50 | 0,68613139 | 0,67153285 | 0,67153285 | 0,67883212 | 0,67737226 | 0,67737226 | 0,67737226 | 0,67737226 | 0,67737226 | 0,67737226 | 0,67737226 | 0,67737226 | 0,67737226 |

*Tabla 68. Scores para la asignatura de Materiales en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,67737226 | 0,67007299 | 0,67883212 | 0,66861314 | 0,66569343 | 0,68175182 | 0,64379562 | 0,65547445 | 0,65839416 | 0,64379562 | 0,64525547 | 0,64379562 | 0,64233577 |
| 2 | 0,67737226 | 0,67007299 | 0,67883212 | 0,66861314 | 0,66277372 | 0,68467153 | 0,64817518 | 0,64233577 | 0,6350365 | 0,65255474 | 0,64087591 | 0,6540146 | 0,6540146 |
| 3 | 0,67737226 | 0,67007299 | 0,67883212 | 0,66861314 | 0,66861314 | 0,6919708 | 0,66423358 | 0,65839416 | 0,6379562 | 0,64963504 | 0,64525547 | 0,65255474 | 0,66131387 |
| 5 | 0,67737226 | 0,67007299 | 0,67883212 | 0,66569343 | 0,66277372 | 0,66277372 | 0,65985401 | 0,65109489 | 0,65255474 | 0,65255474 | 0,65109489 | 0,65839416 | 0,65985401 |
| 7 | 0,67737226 | 0,67007299 | 0,67883212 | 0,67153285 | 0,67007299 | 0,65547445 | 0,6379562 | 0,65693431 | 0,66277372 | 0,65985401 | 0,65547445 | 0,66277372 | 0,66131387 |
| 10 | 0,67737226 | 0,67007299 | 0,67883212 | 0,67007299 | 0,68029197 | 0,66131387 | 0,64525547 | 0,64233577 | 0,64525547 | 0,64525547 | 0,64525547 | 0,64525547 | 0,64525547 |
| 15 | 0,67737226 | 0,67007299 | 0,67883212 | 0,67445255 | 0,6919708 | 0,65985401 | 0,6540146 | 0,64963504 | 0,6540146 | 0,6540146 | 0,64963504 | 0,6540146 | 0,64963504 |
| 20 | 0,67737226 | 0,67007299 | 0,67883212 | 0,70948905 | 0,71970803 | 0,70072993 | 0,69343066 | 0,69635036 | 0,69635036 | 0,69343066 | 0,69343066 | 0,69635036 | 0,69343066 |
| 25 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71824818 | 0,70218978 | 0,70364964 | 0,70364964 | 0,70364964 | 0,70364964 | 0,70364964 | 0,70364964 | 0,70364964 |
| 30 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71678832 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 | 0,70510949 |
| 35 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71678832 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 | 0,7080292 |
| 40 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 | 0,71532847 |
| 45 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71970803 | 0,71970803 | 0,71970803 | 0,71970803 | 0,71970803 | 0,71970803 | 0,71970803 | 0,71970803 | 0,71970803 |
| 50 | 0,67737226 | 0,67007299 | 0,67883212 | 0,71386861 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 | 0,71678832 |

*Tabla 69. Scores para la asignatura de Edos en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,81313869 | 0,76350365 | 0,77956204 | 0,77080292 | 0,7810219 | 0,75182482 | 0,72992701 | 0,70072993 | 0,71094891 | 0,71094891 | 0,71386861 | 0,70948905 | 0,71678832 |
| 2 | 0,81313869 | 0,76350365 | 0,77956204 | 0,7649635 | 0,77518248 | 0,75474453 | 0,70948905 | 0,6919708 | 0,71532847 | 0,71678832 | 0,70656934 | 0,70072993 | 0,70948905 |
| 3 | 0,81313869 | 0,76350365 | 0,77956204 | 0,76350365 | 0,77664234 | 0,75182482 | 0,73576642 | 0,71240876 | 0,71240876 | 0,71678832 | 0,71532847 | 0,72554745 | 0,71386861 |
| 5 | 0,81313869 | 0,76350365 | 0,77956204 | 0,77226277 | 0,76058394 | 0,7649635 | 0,7459854 | 0,74306569 | 0,73284672 | 0,73138686 | 0,73868613 | 0,74452555 | 0,73576642 |
| 7 | 0,81313869 | 0,76350365 | 0,77956204 | 0,77810219 | 0,77372263 | 0,78978102 | 0,76058394 | 0,75474453 | 0,77226277 | 0,76934307 | 0,76350365 | 0,76934307 | 0,75912409 |
| 10 | 0,81313869 | 0,76350365 | 0,77956204 | 0,77810219 | 0,7810219 | 0,77518248 | 0,75182482 | 0,75620438 | 0,75474453 | 0,75474453 | 0,75328467 | 0,75182482 | 0,75182482 |
| 15 | 0,81313869 | 0,76350365 | 0,77956204 | 0,77080292 | 0,78540146 | 0,7810219 | 0,7810219 | 0,7810219 | 0,7810219 | 0,7810219 | 0,7810219 | 0,7810219 | 0,7810219 |
| 20 | 0,81313869 | 0,76350365 | 0,77956204 | 0,7620438 | 0,77664234 | 0,77080292 | 0,77080292 | 0,77518248 | 0,77518248 | 0,77518248 | 0,77080292 | 0,77080292 | 0,77080292 |
| 25 | 0,81313869 | 0,76350365 | 0,80145985 | 0,79270073 | 0,79416058 | 0,79416058 | 0,79416058 | 0,79416058 | 0,79416058 | 0,79416058 | 0,79416058 | 0,79416058 | 0,79416058 |
| 30 | 0,81313869 | 0,76350365 | 0,80145985 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 |
| 35 | 0,81313869 | 0,76350365 | 0,80145985 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 |
| 40 | 0,81313869 | 0,76350365 | 0,80145985 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 |
| 45 | 0,81313869 | 0,76350365 | 0,80145985 | 0,79270073 | 0,79270073 | 0,79270073 | 0,79270073 | 0,79270073 | 0,79270073 | 0,79270073 | 0,79270073 | 0,79270073 | 0,79270073 |
| 50 | 0,81313869 | 0,76350365 | 0,80145985 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 | 0,78978102 |

*Tabla 70. Scores para la asignatura de Informática en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,63941606 | 0,6540146 | 0,65985401 | 0,67153285 | 0,64963504 | 0,64817518 | 0,65693431 | 0,64963504 | 0,64671533 | 0,6350365 | 0,63649635 | 0,65693431 | 0,63211679 |
| 2 | 0,63941606 | 0,6540146 | 0,65985401 | 0,67153285 | 0,65109489 | 0,6540146 | 0,63357664 | 0,64963504 | 0,64525547 | 0,62335766 | 0,65109489 | 0,64379562 | 0,64233577 |
| 3 | 0,63941606 | 0,6540146 | 0,65985401 | 0,6729927 | 0,65255474 | 0,64671533 | 0,63649635 | 0,62919708 | 0,62773723 | 0,64233577 | 0,64233577 | 0,64087591 | 0,65109489 |
| 5 | 0,63941606 | 0,6540146 | 0,65985401 | 0,67445255 | 0,65693431 | 0,62335766 | 0,62919708 | 0,62919708 | 0,6350365 | 0,6379562 | 0,62773723 | 0,63941606 | 0,62919708 |
| 7 | 0,63941606 | 0,6540146 | 0,65985401 | 0,67737226 | 0,65985401 | 0,64087591 | 0,6189781 | 0,62627737 | 0,61751825 | 0,62335766 | 0,62627737 | 0,62189781 | 0,62043796 |
| 10 | 0,63941606 | 0,6540146 | 0,65985401 | 0,67737226 | 0,65693431 | 0,63649635 | 0,62043796 | 0,6189781 | 0,62335766 | 0,6189781 | 0,62043796 | 0,6189781 | 0,62189781 |
| 15 | 0,63941606 | 0,6540146 | 0,65985401 | 0,67591241 | 0,66861314 | 0,64671533 | 0,63941606 | 0,63941606 | 0,64087591 | 0,64087591 | 0,64087591 | 0,64087591 | 0,64087591 |
| 20 | 0,63941606 | 0,6540146 | 0,65985401 | 0,67591241 | 0,66715328 | 0,66423358 | 0,65839416 | 0,65693431 | 0,65693431 | 0,65693431 | 0,65839416 | 0,65693431 | 0,65693431 |
| 25 | 0,63941606 | 0,6540146 | 0,65985401 | 0,6540146 | 0,64817518 | 0,64817518 | 0,64087591 | 0,64087591 | 0,64087591 | 0,64087591 | 0,64087591 | 0,64087591 | 0,64087591 |
| 30 | 0,63941606 | 0,6540146 | 0,65985401 | 0,64963504 | 0,64963504 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 |
| 35 | 0,63941606 | 0,6540146 | 0,65985401 | 0,64963504 | 0,65255474 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 |
| 40 | 0,63941606 | 0,6540146 | 0,65985401 | 0,64817518 | 0,65255474 | 0,65255474 | 0,65255474 | 0,65255474 | 0,65255474 | 0,65255474 | 0,65255474 | 0,65255474 | 0,65255474 |
| 45 | 0,63941606 | 0,6540146 | 0,65985401 | 0,64525547 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 | 0,64817518 |
| 50 | 0,63941606 | 0,6540146 | 0,65985401 | 0,66423358 | 0,66423358 | 0,66423358 | 0,66423358 | 0,66423358 | 0,66423358 | 0,66423358 | 0,66423358 | 0,66423358 | 0,66423358 |

*Tabla 71. Scores para la asignatura de Mecánica en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0,48868778 | 0,54298643 | 0,57918552 | 0,56561086 | 0,46153846 | 0,47963801 | 0,48868778 | 0,47963801 | 0,49773756 | 0,48868778 | 0,47963801 | 0,49773756 |
| 2 | 0 | 0,48868778 | 0,54298643 | 0,58823529 | 0,57013575 | 0,47963801 | 0,5158371 | 0,50226244 | 0,50678733 | 0,5158371 | 0,50678733 | 0,51131222 | 0,50678733 |
| 3 | 0 | 0,48868778 | 0,54298643 | 0,58823529 | 0,56108597 | 0,4841629 | 0,53846154 | 0,47963801 | 0,50678733 | 0,51131222 | 0,50226244 | 0,49321267 | 0,47511312 |
| 5 | 0 | 0,48868778 | 0,54298643 | 0,58371041 | 0,53846154 | 0,55656109 | 0,49321267 | 0,49773756 | 0,47511312 | 0,5158371 | 0,5158371 | 0,51131222 | 0,49321267 |
| 7 | 0 | 0,48868778 | 0,53846154 | 0,59276018 | 0,59728507 | 0,53393665 | 0,46153846 | 0,4841629 | 0,4841629 | 0,47963801 | 0,47511312 | 0,47511312 | 0,49321267 |
| 10 | 0 | 0,48868778 | 0,53846154 | 0,57466063 | 0,40271493 | 0,42533937 | 0,46606335 | 0,46606335 | 0,46606335 | 0,46606335 | 0,47058824 | 0,47058824 | 0,46153846 |
| 15 | 0 | 0,48868778 | 0,53846154 | 0,52036199 | 0,36651584 | 0,4479638 | 0,44343891 | 0,44343891 | 0,44343891 | 0,44343891 | 0,4479638 | 0,44343891 | 0,44343891 |
| 20 | 0 | 0,48868778 | 0,53846154 | 0,35294118 | 0,45701357 | 0,46606335 | 0,49773756 | 0,47963801 | 0,49773756 | 0,47963801 | 0,49773756 | 0,47963801 | 0,49773756 |
| 25 | 0 | 0,48868778 | 0,53846154 | 0,33936652 | 0,47963801 | 0,49773756 | 0,51131222 | 0,51131222 | 0,51131222 | 0,51131222 | 0,51131222 | 0,51131222 | 0,51131222 |
| 30 | 0 | 0,48868778 | 0,53846154 | 0,32126697 | 0,42986425 | 0,41628959 | 0,41628959 | 0,41628959 | 0,41628959 | 0,41628959 | 0,41628959 | 0,41628959 | 0,41628959 |
| 35 | 0 | 0,48868778 | 0,53846154 | 0,32126697 | 0,42986425 | 0,46153846 | 0,46153846 | 0,46153846 | 0,46153846 | 0,46153846 | 0,46153846 | 0,46153846 | 0,46153846 |
| 40 | 0 | 0,48868778 | 0,53846154 | 0,32126697 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 |
| 45 | 0 | 0,48868778 | 0,53846154 | 0,32126697 | 0,42081448 | 0,42081448 | 0,42081448 | 0,42081448 | 0,42081448 | 0,42081448 | 0,42081448 | 0,42081448 | 0,42081448 |
| 50 | 0 | 0,48868778 | 0,53846154 | 0,32126697 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 | 0,45248869 |

*Tabla 72. Recall del suspenso para la asignatura de Electromagnetismo en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0,02197802 | 0,04395604 | 0,05494505 | 0,20879121 | 0,27472527 | 0,26373626 | 0,31868132 | 0,28571429 | 0,2967033 | 0,32967033 | 0,28571429 | 0,31868132 |
| 2 | 0 | 0,02197802 | 0,04395604 | 0,05494505 | 0,20879121 | 0,27472527 | 0,28571429 | 0,31868132 | 0,32967033 | 0,30769231 | 0,2967033 | 0,30769231 | 0,31868132 |
| 3 | 0 | 0,02197802 | 0,04395604 | 0,05494505 | 0,20879121 | 0,27472527 | 0,2967033 | 0,30769231 | 0,2967033 | 0,28571429 | 0,30769231 | 0,28571429 | 0,31868132 |
| 5 | 0 | 0,02197802 | 0,04395604 | 0,07692308 | 0,20879121 | 0,24175824 | 0,26373626 | 0,26373626 | 0,30769231 | 0,2967033 | 0,26373626 | 0,2967033 | 0,27472527 |
| 7 | 0 | 0,02197802 | 0,02197802 | 0,0989011 | 0,17582418 | 0,1978022 | 0,21978022 | 0,21978022 | 0,21978022 | 0,21978022 | 0,21978022 | 0,21978022 | 0,21978022 |
| 10 | 0 | 0,17582418 | 0,10989011 | 0,20879121 | 0,26373626 | 0,30769231 | 0,32967033 | 0,32967033 | 0,32967033 | 0,31868132 | 0,31868132 | 0,32967033 | 0,31868132 |
| 15 | 0 | 0,17582418 | 0,10989011 | 0,18681319 | 0,15384615 | 0,15384615 | 0,15384615 | 0,15384615 | 0,15384615 | 0,15384615 | 0,15384615 | 0,15384615 | 0,15384615 |
| 20 | 0 | 0,17582418 | 0,08791209 | 0,23076923 | 0,23076923 | 0,23076923 | 0,23076923 | 0,23076923 | 0,23076923 | 0,23076923 | 0,23076923 | 0,23076923 | 0,23076923 |
| 25 | 0 | 0,17582418 | 0,08791209 | 0,16483516 | 0,16483516 | 0,16483516 | 0,16483516 | 0,16483516 | 0,16483516 | 0,16483516 | 0,16483516 | 0,16483516 | 0,16483516 |
| 30 | 0 | 0,17582418 | 0,12087912 | 0,1978022 | 0,1978022 | 0,1978022 | 0,1978022 | 0,1978022 | 0,1978022 | 0,1978022 | 0,1978022 | 0,1978022 | 0,1978022 |
| 35 | 0 | 0,17582418 | 0,14285714 | 0,18681319 | 0,18681319 | 0,18681319 | 0,18681319 | 0,18681319 | 0,18681319 | 0,18681319 | 0,18681319 | 0,18681319 | 0,18681319 |
| 40 | 0 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 |
| 45 | 0 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 |
| 50 | 0 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 | 0,17582418 |

*Tabla 73. Recall del suspenso para la asignatura de Métodos en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0,16744186 | 0,16744186 | 0,40465116 | 0,31627907 | 0,32093023 | 0,36744186 | 0,37209302 | 0,41860465 | 0,36744186 | 0,40930233 | 0,41860465 | 0,35813953 |
| 2 | 0 | 0,16744186 | 0,16744186 | 0,40465116 | 0,30697674 | 0,34418605 | 0,4 | 0,41860465 | 0,43255814 | 0,38604651 | 0,40930233 | 0,42790698 | 0,4372093 |
| 3 | 0 | 0,16744186 | 0,16744186 | 0,4 | 0,3255814 | 0,38139535 | 0,43255814 | 0,43255814 | 0,41860465 | 0,4372093 | 0,40930233 | 0,4 | 0,39534884 |
| 5 | 0 | 0,16744186 | 0,16744186 | 0,39534884 | 0,33488372 | 0,35348837 | 0,39534884 | 0,39534884 | 0,4 | 0,40465116 | 0,38604651 | 0,38139535 | 0,37209302 |
| 7 | 0 | 0,16744186 | 0,16744186 | 0,39534884 | 0,35348837 | 0,38604651 | 0,38604651 | 0,39534884 | 0,4 | 0,39534884 | 0,40465116 | 0,39069767 | 0,4 |
| 10 | 0 | 0,16744186 | 0,16744186 | 0,39534884 | 0,33488372 | 0,38604651 | 0,38139535 | 0,38139535 | 0,38139535 | 0,38139535 | 0,38139535 | 0,38139535 | 0,38139535 |
| 15 | 0 | 0,16744186 | 0,16744186 | 0,39069767 | 0,38139535 | 0,40465116 | 0,40930233 | 0,40465116 | 0,40465116 | 0,40465116 | 0,40465116 | 0,40465116 | 0,40465116 |
| 20 | 0 | 0,16744186 | 0,16744186 | 0,40930233 | 0,38139535 | 0,39534884 | 0,39534884 | 0,39534884 | 0,39534884 | 0,39534884 | 0,39534884 | 0,39534884 | 0,39534884 |
| 25 | 0 | 0,16744186 | 0,16744186 | 0,42325581 | 0,37209302 | 0,28837209 | 0,28837209 | 0,28837209 | 0,28837209 | 0,28837209 | 0,28837209 | 0,28837209 | 0,28837209 |
| 30 | 0 | 0,16744186 | 0,16744186 | 0,41395349 | 0,35813953 | 0,2744186 | 0,2744186 | 0,2744186 | 0,2744186 | 0,2744186 | 0,2744186 | 0,2744186 | 0,2744186 |
| 35 | 0 | 0,16744186 | 0,16744186 | 0,4 | 0,26511628 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 |
| 40 | 0 | 0,16744186 | 0,16744186 | 0,4 | 0,26511628 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 | 0,21395349 |
| 45 | 0 | 0,16744186 | 0,16744186 | 0,4 | 0,26511628 | 0,2372093 | 0,2372093 | 0,2372093 | 0,2372093 | 0,2372093 | 0,2372093 | 0,2372093 | 0,2372093 |
| 50 | 0 | 0,16744186 | 0,16744186 | 0,4372093 | 0,30232558 | 0,30232558 | 0,30232558 | 0,30232558 | 0,30232558 | 0,30232558 | 0,30232558 | 0,30232558 | 0,30232558 |

*Tabla 74. Recall del suspenso para la asignatura de Materiales en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0,24647887 | 0,11971831 | 0,1056338 | 0,17605634 | 0,28169014 | 0,28169014 | 0,27464789 | 0,30985915 | 0,28873239 | 0,29577465 | 0,34507042 |
| 2 | 0 | 0 | 0,24647887 | 0,11971831 | 0,14084507 | 0,20422535 | 0,28169014 | 0,27464789 | 0,3028169 | 0,32394366 | 0,28873239 | 0,32394366 | 0,33098592 |
| 3 | 0 | 0 | 0,24647887 | 0,09859155 | 0,11267606 | 0,16197183 | 0,28169014 | 0,29577465 | 0,30985915 | 0,31690141 | 0,30985915 | 0,29577465 | 0,32394366 |
| 5 | 0 | 0 | 0,24647887 | 0,11267606 | 0,17605634 | 0,26056338 | 0,32394366 | 0,33802817 | 0,31690141 | 0,3028169 | 0,33098592 | 0,35211268 | 0,33802817 |
| 7 | 0 | 0 | 0,22535211 | 0,11971831 | 0,22535211 | 0,28169014 | 0,26760563 | 0,28169014 | 0,27464789 | 0,27464789 | 0,28873239 | 0,28873239 | 0,26056338 |
| 10 | 0 | 0 | 0,22535211 | 0,11971831 | 0,23943662 | 0,29577465 | 0,34507042 | 0,34507042 | 0,33802817 | 0,33802817 | 0,34507042 | 0,34507042 | 0,33802817 |
| 15 | 0 | 0 | 0,21126761 | 0,09859155 | 0,15492958 | 0,25352113 | 0,26760563 | 0,26760563 | 0,26760563 | 0,26056338 | 0,26760563 | 0,26056338 | 0,26056338 |
| 20 | 0 | 0 | 0,21126761 | 0,07746479 | 0,22535211 | 0,22535211 | 0,22535211 | 0,22535211 | 0,22535211 | 0,22535211 | 0,22535211 | 0,22535211 | 0,22535211 |
| 25 | 0 | 0 | 0,21126761 | 0,0915493 | 0,1056338 | 0,1056338 | 0,1056338 | 0,1056338 | 0,1056338 | 0,1056338 | 0,1056338 | 0,1056338 | 0,1056338 |
| 30 | 0 | 0 | 0,21126761 | 0,0915493 | 0,12676056 | 0,12676056 | 0,12676056 | 0,12676056 | 0,12676056 | 0,12676056 | 0,12676056 | 0,12676056 | 0,12676056 |
| 35 | 0 | 0 | 0,21126761 | 0,11267606 | 0,11267606 | 0,11267606 | 0,11267606 | 0,11267606 | 0,11267606 | 0,11267606 | 0,11267606 | 0,11267606 | 0,11267606 |
| 40 | 0 | 0 | 0,21126761 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 |
| 45 | 0 | 0 | 0,21126761 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 |
| 50 | 0 | 0 | 0,21126761 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 | 0,16197183 |

*Tabla 75. Recall del suspenso para la asignatura de Edos en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0,25 | 0,2421875 | 0,3046875 | 0,2265625 | 0,265625 | 0,390625 | 0,390625 | 0,3984375 | 0,390625 | 0,34375 | 0,3671875 | 0,3984375 |
| 2 | 0 | 0,25 | 0,2421875 | 0,3046875 | 0,2265625 | 0,28125 | 0,3984375 | 0,40625 | 0,4296875 | 0,40625 | 0,421875 | 0,4296875 | 0,390625 |
| 3 | 0 | 0,25 | 0,2421875 | 0,3046875 | 0,2109375 | 0,296875 | 0,390625 | 0,3984375 | 0,390625 | 0,390625 | 0,3828125 | 0,3828125 | 0,3984375 |
| 5 | 0 | 0,25 | 0,2421875 | 0,296875 | 0,2890625 | 0,359375 | 0,3515625 | 0,3203125 | 0,359375 | 0,3515625 | 0,328125 | 0,3359375 | 0,3515625 |
| 7 | 0 | 0,25 | 0,2421875 | 0,296875 | 0,296875 | 0,328125 | 0,359375 | 0,3828125 | 0,359375 | 0,390625 | 0,3828125 | 0,3515625 | 0,34375 |
| 10 | 0 | 0,25 | 0,2421875 | 0,3046875 | 0,2734375 | 0,3359375 | 0,359375 | 0,3515625 | 0,3515625 | 0,359375 | 0,3515625 | 0,359375 | 0,3671875 |
| 15 | 0 | 0,25 | 0,2421875 | 0,3046875 | 0,2734375 | 0,328125 | 0,328125 | 0,328125 | 0,328125 | 0,328125 | 0,328125 | 0,328125 | 0,328125 |
| 20 | 0 | 0,25 | 0,2421875 | 0,3515625 | 0,2734375 | 0,3125 | 0,3125 | 0,28125 | 0,28125 | 0,3125 | 0,3125 | 0,28125 | 0,28125 |
| 25 | 0 | 0,25 | 0,0859375 | 0,2578125 | 0,28125 | 0,28125 | 0,28125 | 0,28125 | 0,28125 | 0,28125 | 0,28125 | 0,28125 | 0,28125 |
| 30 | 0 | 0,25 | 0,0859375 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 |
| 35 | 0 | 0,25 | 0,0859375 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 |
| 40 | 0 | 0,25 | 0,0859375 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 | 0,2890625 |
| 45 | 0 | 0,25 | 0,0859375 | 0,3203125 | 0,3203125 | 0,3203125 | 0,3203125 | 0,3203125 | 0,3203125 | 0,3203125 | 0,3203125 | 0,3203125 | 0,3203125 |
| 50 | 0 | 0,25 | 0,0859375 | 0,25 | 0,25 | 0,25 | 0,25 | 0,25 | 0,25 | 0,25 | 0,25 | 0,25 | 0,25 |

*Tabla 76. Recall del suspenso para la asignatura de Informática en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,83591331 | 0,66873065 | 0,78328173 | 0,74922601 | 0,6873065 | 0,74613003 | 0,71517028 | 0,64705882 | 0,65325077 | 0,63157895 | 0,6501548 | 0,64705882 | 0,6377709 |
| 2 | 0,83591331 | 0,66873065 | 0,78328173 | 0,74922601 | 0,69040248 | 0,75232198 | 0,73065015 | 0,6749226 | 0,6625387 | 0,66873065 | 0,68111455 | 0,67801858 | 0,66873065 |
| 3 | 0,83591331 | 0,66873065 | 0,78328173 | 0,74922601 | 0,6873065 | 0,73065015 | 0,70897833 | 0,65325077 | 0,6377709 | 0,64396285 | 0,64396285 | 0,65634675 | 0,64705882 |
| 5 | 0,83591331 | 0,66873065 | 0,78328173 | 0,74922601 | 0,74613003 | 0,73993808 | 0,6501548 | 0,6501548 | 0,64396285 | 0,65944272 | 0,66873065 | 0,65325077 | 0,67182663 |
| 7 | 0,83591331 | 0,66873065 | 0,78328173 | 0,74922601 | 0,6996904 | 0,69040248 | 0,63467492 | 0,625387 | 0,625387 | 0,625387 | 0,63157895 | 0,625387 | 0,62848297 |
| 10 | 0,83591331 | 0,66873065 | 0,78328173 | 0,76160991 | 0,68421053 | 0,6749226 | 0,64086687 | 0,64086687 | 0,64086687 | 0,6377709 | 0,64086687 | 0,64086687 | 0,64086687 |
| 15 | 0,83591331 | 0,66873065 | 0,78328173 | 0,76780186 | 0,72136223 | 0,73065015 | 0,7244582 | 0,73684211 | 0,73374613 | 0,73374613 | 0,73374613 | 0,72136223 | 0,7244582 |
| 20 | 0,83591331 | 0,66873065 | 0,78328173 | 0,74922601 | 0,72136223 | 0,73684211 | 0,7120743 | 0,7120743 | 0,7244582 | 0,7244582 | 0,7120743 | 0,7244582 | 0,7244582 |
| 25 | 0,83591331 | 0,66873065 | 0,78328173 | 0,74922601 | 0,6873065 | 0,65944272 | 0,64705882 | 0,64705882 | 0,64705882 | 0,64705882 | 0,64705882 | 0,64705882 | 0,64705882 |
| 30 | 0,83591331 | 0,66873065 | 0,78328173 | 0,78328173 | 0,7244582 | 0,6996904 | 0,6996904 | 0,6996904 | 0,6996904 | 0,6996904 | 0,6996904 | 0,6996904 | 0,6996904 |
| 35 | 0,83591331 | 0,66873065 | 0,78328173 | 0,74303406 | 0,75232198 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 |
| 40 | 0,83591331 | 0,66873065 | 0,78328173 | 0,74613003 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 | 0,72755418 |
| 45 | 0,83591331 | 0,66873065 | 0,78328173 | 0,77089783 | 0,74922601 | 0,74922601 | 0,74922601 | 0,74922601 | 0,74922601 | 0,74922601 | 0,74922601 | 0,74922601 | 0,74922601 |
| 50 | 0,83591331 | 0,66873065 | 0,78328173 | 0,75232198 | 0,75232198 | 0,75232198 | 0,75232198 | 0,75232198 | 0,75232198 | 0,75232198 | 0,75232198 | 0,75232198 | 0,75232198 |

*Tabla 77. Recall del suspenso para la asignatura de Mecánica en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0,75646552 | 0,74353448 | 0,70905172 | 0,71551724 | 0,78663793 | 0,73491379 | 0,71982759 | 0,71551724 | 0,73060345 | 0,71982759 | 0,71767241 | 0,72844828 |
| 2 | 1 | 0,75646552 | 0,74353448 | 0,70689655 | 0,7112069 | 0,78232759 | 0,73060345 | 0,71336207 | 0,71551724 | 0,72844828 | 0,71982759 | 0,72413793 | 0,71336207 |
| 3 | 1 | 0,75646552 | 0,74353448 | 0,70689655 | 0,71982759 | 0,79094828 | 0,72844828 | 0,72844828 | 0,72198276 | 0,72198276 | 0,71336207 | 0,7262931 | 0,73060345 |
| 5 | 1 | 0,75646552 | 0,74353448 | 0,70474138 | 0,72198276 | 0,71336207 | 0,73491379 | 0,73275862 | 0,73275862 | 0,73491379 | 0,73275862 | 0,73275862 | 0,74568966 |
| 7 | 1 | 0,75646552 | 0,74568966 | 0,70905172 | 0,70474138 | 0,71336207 | 0,73491379 | 0,74353448 | 0,73922414 | 0,73922414 | 0,73922414 | 0,74353448 | 0,74137931 |
| 10 | 1 | 0,75646552 | 0,74568966 | 0,71551724 | 0,8125 | 0,77586207 | 0,7262931 | 0,7262931 | 0,73060345 | 0,7262931 | 0,72413793 | 0,72413793 | 0,73275862 |
| 15 | 1 | 0,75646552 | 0,74568966 | 0,74784483 | 0,84698276 | 0,75431034 | 0,74784483 | 0,74784483 | 0,74784483 | 0,74784483 | 0,75215517 | 0,74784483 | 0,74784483 |
| 20 | 1 | 0,75646552 | 0,74568966 | 0,87931034 | 0,84913793 | 0,80818966 | 0,79094828 | 0,79525862 | 0,79094828 | 0,79525862 | 0,79094828 | 0,79525862 | 0,79094828 |
| 25 | 1 | 0,75646552 | 0,74568966 | 0,89224138 | 0,83189655 | 0,79956897 | 0,79525862 | 0,79525862 | 0,79525862 | 0,79525862 | 0,79525862 | 0,79525862 | 0,79525862 |
| 30 | 1 | 0,75646552 | 0,74568966 | 0,90086207 | 0,85344828 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 |
| 35 | 1 | 0,75646552 | 0,74568966 | 0,90086207 | 0,85344828 | 0,82543103 | 0,82543103 | 0,82543103 | 0,82543103 | 0,82543103 | 0,82543103 | 0,82543103 | 0,82543103 |
| 40 | 1 | 0,75646552 | 0,74568966 | 0,90086207 | 0,84051724 | 0,84051724 | 0,84051724 | 0,84051724 | 0,84051724 | 0,84051724 | 0,84051724 | 0,84051724 | 0,84051724 |
| 45 | 1 | 0,75646552 | 0,74568966 | 0,90086207 | 0,86206897 | 0,86206897 | 0,86206897 | 0,86206897 | 0,86206897 | 0,86206897 | 0,86206897 | 0,86206897 | 0,86206897 |
| 50 | 1 | 0,75646552 | 0,74568966 | 0,90086207 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 | 0,84267241 |

*Tabla 78. Recall del aprobado para la asignatura de Electromagnetismo en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0,99494949 | 0,99494949 | 0,98484848 | 0,94107744 | 0,91245791 | 0,84343434 | 0,85185185 | 0,86363636 | 0,85353535 | 0,87037037 | 0,86363636 | 0,85185185 |
| 2 | 1 | 0,99494949 | 0,99494949 | 0,98653199 | 0,93939394 | 0,9040404 | 0,83333333 | 0,83670034 | 0,83501684 | 0,84343434 | 0,84343434 | 0,84511785 | 0,83501684 |
| 3 | 1 | 0,99494949 | 0,99494949 | 0,99326599 | 0,94444444 | 0,9023569 | 0,85521886 | 0,85016835 | 0,86868687 | 0,86195286 | 0,85016835 | 0,87373737 | 0,84848485 |
| 5 | 1 | 0,99494949 | 0,99494949 | 0,98148148 | 0,94107744 | 0,92592593 | 0,87542088 | 0,87542088 | 0,86868687 | 0,88215488 | 0,87205387 | 0,87542088 | 0,87205387 |
| 7 | 1 | 0,99494949 | 0,99494949 | 0,97979798 | 0,95791246 | 0,94107744 | 0,9040404 | 0,91077441 | 0,90572391 | 0,91414141 | 0,90740741 | 0,91077441 | 0,9040404 |
| 10 | 1 | 0,98316498 | 0,97811448 | 0,96969697 | 0,95117845 | 0,92592593 | 0,89393939 | 0,89393939 | 0,89393939 | 0,8989899 | 0,8989899 | 0,89393939 | 0,8989899 |
| 15 | 1 | 0,98316498 | 0,97811448 | 0,95286195 | 0,96801347 | 0,95622896 | 0,95622896 | 0,95622896 | 0,95622896 | 0,95622896 | 0,95622896 | 0,95622896 | 0,95622896 |
| 20 | 1 | 0,98316498 | 0,99158249 | 0,96296296 | 0,96296296 | 0,96296296 | 0,96296296 | 0,96296296 | 0,96296296 | 0,96296296 | 0,96296296 | 0,96296296 | 0,96296296 |
| 25 | 1 | 0,98316498 | 0,99158249 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 |
| 30 | 1 | 0,98316498 | 0,99158249 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 | 0,96632997 |
| 35 | 1 | 0,98316498 | 0,98989899 | 0,96969697 | 0,96969697 | 0,96969697 | 0,96969697 | 0,96969697 | 0,96969697 | 0,96969697 | 0,96969697 | 0,96969697 | 0,96969697 |
| 40 | 1 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 |
| 45 | 1 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 |
| 50 | 1 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 | 0,98316498 |

*Tabla 79. Recall del aprobado para la asignatura de Métodos en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0,90212766 | 0,90212766 | 0,80212766 | 0,84893617 | 0,76808511 | 0,74680851 | 0,75319149 | 0,72553191 | 0,72978723 | 0,73404255 | 0,73829787 | 0,74042553 |
| 2 | 1 | 0,90212766 | 0,90212766 | 0,80425532 | 0,85319149 | 0,76808511 | 0,73829787 | 0,73617021 | 0,72765957 | 0,73404255 | 0,7212766 | 0,7212766 | 0,71489362 |
| 3 | 1 | 0,90212766 | 0,90212766 | 0,80425532 | 0,84893617 | 0,75744681 | 0,75531915 | 0,74680851 | 0,75106383 | 0,73829787 | 0,73829787 | 0,75744681 | 0,75106383 |
| 5 | 1 | 0,90212766 | 0,90212766 | 0,80212766 | 0,8212766 | 0,76382979 | 0,75744681 | 0,76382979 | 0,77021277 | 0,75531915 | 0,77446809 | 0,76595745 | 0,76808511 |
| 7 | 1 | 0,90212766 | 0,90212766 | 0,80212766 | 0,81914894 | 0,75957447 | 0,74255319 | 0,75106383 | 0,74468085 | 0,75106383 | 0,75319149 | 0,75319149 | 0,75106383 |
| 10 | 1 | 0,90212766 | 0,90212766 | 0,80212766 | 0,83829787 | 0,74042553 | 0,74255319 | 0,74893617 | 0,74255319 | 0,74893617 | 0,74893617 | 0,74255319 | 0,74893617 |
| 15 | 1 | 0,90212766 | 0,90212766 | 0,79787234 | 0,78085106 | 0,76170213 | 0,76170213 | 0,76170213 | 0,76170213 | 0,76170213 | 0,76170213 | 0,76170213 | 0,76170213 |
| 20 | 1 | 0,90212766 | 0,90212766 | 0,7787234 | 0,78085106 | 0,78085106 | 0,78085106 | 0,78085106 | 0,78085106 | 0,78085106 | 0,78085106 | 0,78085106 | 0,78085106 |
| 25 | 1 | 0,90212766 | 0,90212766 | 0,80851064 | 0,81914894 | 0,85744681 | 0,85744681 | 0,85744681 | 0,85744681 | 0,85744681 | 0,85744681 | 0,85744681 | 0,85744681 |
| 30 | 1 | 0,90212766 | 0,90212766 | 0,82340426 | 0,83404255 | 0,87234043 | 0,87234043 | 0,87234043 | 0,87234043 | 0,87234043 | 0,87234043 | 0,87234043 | 0,87234043 |
| 35 | 1 | 0,90212766 | 0,90212766 | 0,81702128 | 0,87659574 | 0,9 | 0,9 | 0,9 | 0,9 | 0,9 | 0,9 | 0,9 | 0,9 |
| 40 | 1 | 0,90212766 | 0,90212766 | 0,81702128 | 0,87659574 | 0,9 | 0,9 | 0,9 | 0,9 | 0,9 | 0,9 | 0,9 | 0,9 |
| 45 | 1 | 0,90212766 | 0,90212766 | 0,81702128 | 0,87659574 | 0,90212766 | 0,90212766 | 0,90212766 | 0,90212766 | 0,90212766 | 0,90212766 | 0,90212766 | 0,90212766 |
| 50 | 1 | 0,90212766 | 0,90212766 | 0,7893617 | 0,84893617 | 0,84893617 | 0,84893617 | 0,84893617 | 0,84893617 | 0,84893617 | 0,84893617 | 0,84893617 | 0,84893617 |

*Tabla 80. Recall del aprobado para la asignatura de Materiales en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0,93370166 | 0,95948435 | 0,94475138 | 0,90791897 | 0,86003683 | 0,84530387 | 0,83425414 | 0,83977901 | 0,84714549 | 0,83425414 | 0,85635359 |
| 2 | 1 | 1 | 0,93370166 | 0,95948435 | 0,93554328 | 0,88766114 | 0,84346225 | 0,83241252 | 0,80294659 | 0,81767956 | 0,82688766 | 0,81583794 | 0,8121547 |
| 3 | 1 | 1 | 0,93370166 | 0,95764273 | 0,93370166 | 0,88766114 | 0,84162063 | 0,81583794 | 0,82872928 | 0,82872928 | 0,83793738 | 0,82504604 | 0,8121547 |
| 5 | 1 | 1 | 0,93370166 | 0,95211786 | 0,93370166 | 0,90055249 | 0,83425414 | 0,8305709 | 0,82504604 | 0,83425414 | 0,83425414 | 0,83793738 | 0,83609576 |
| 7 | 1 | 1 | 0,93554328 | 0,94843462 | 0,87661142 | 0,88029466 | 0,84898711 | 0,85267035 | 0,84714549 | 0,85082873 | 0,84898711 | 0,85082873 | 0,84898711 |
| 10 | 1 | 1 | 0,93554328 | 0,94843462 | 0,87661142 | 0,87845304 | 0,84162063 | 0,84162063 | 0,83425414 | 0,83425414 | 0,84162063 | 0,84162063 | 0,83425414 |
| 15 | 1 | 1 | 0,93554328 | 0,95027624 | 0,93186004 | 0,8839779 | 0,87661142 | 0,87661142 | 0,87661142 | 0,87108656 | 0,87661142 | 0,87108656 | 0,87108656 |
| 20 | 1 | 1 | 0,93554328 | 0,97790055 | 0,89686924 | 0,90791897 | 0,90791897 | 0,90791897 | 0,90791897 | 0,90791897 | 0,90791897 | 0,90791897 | 0,90791897 |
| 25 | 1 | 1 | 0,93554328 | 0,96132597 | 0,93001842 | 0,93001842 | 0,93001842 | 0,93001842 | 0,93001842 | 0,93001842 | 0,93001842 | 0,93001842 | 0,93001842 |
| 30 | 1 | 1 | 0,93554328 | 0,96316759 | 0,9373849 | 0,9373849 | 0,9373849 | 0,9373849 | 0,9373849 | 0,9373849 | 0,9373849 | 0,9373849 | 0,9373849 |
| 35 | 1 | 1 | 0,93554328 | 0,93554328 | 0,93554328 | 0,93554328 | 0,93554328 | 0,93554328 | 0,93554328 | 0,93554328 | 0,93554328 | 0,93554328 | 0,93554328 |
| 40 | 1 | 1 | 0,93554328 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 |
| 45 | 1 | 1 | 0,93554328 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 |
| 50 | 1 | 1 | 0,93554328 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 | 0,92449355 |

*Tabla 81. Recall del aprobado para la asignatura de Edos en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0,88150808 | 0,90305206 | 0,87612208 | 0,90664273 | 0,87791741 | 0,8043088 | 0,79533214 | 0,78456014 | 0,78096948 | 0,79533214 | 0,78456014 | 0,78994614 |
| 2 | 1 | 0,88150808 | 0,90305206 | 0,87073609 | 0,89766607 | 0,85278276 | 0,78994614 | 0,77917415 | 0,77558348 | 0,78994614 | 0,76660682 | 0,76096948 | 0,78276481 |
| 3 | 1 | 0,88150808 | 0,90305206 | 0,86894075 | 0,8994614 | 0,85278276 | 0,80610413 | 0,78276481 | 0,78815081 | 0,78994614 | 0,79174147 | 0,78815081 | 0,79533214 |
| 5 | 1 | 0,88150808 | 0,90305206 | 0,88150808 | 0,86894075 | 0,86175943 | 0,83662478 | 0,82585278 | 0,83303411 | 0,82764811 | 0,82764811 | 0,81867145 | 0,83123878 |
| 7 | 1 | 0,88150808 | 0,90305206 | 0,88868941 | 0,88330341 | 0,89587074 | 0,85278276 | 0,86355476 | 0,86355476 | 0,86175943 | 0,85996409 | 0,85637343 | 0,87253142 |
| 10 | 1 | 0,88150808 | 0,90305206 | 0,88868941 | 0,89766607 | 0,87073609 | 0,84560144 | 0,8438061 | 0,8438061 | 0,84560144 | 0,8438061 | 0,84021544 | 0,8438061 |
| 15 | 1 | 0,88150808 | 0,90305206 | 0,87791741 | 0,90305206 | 0,88509874 | 0,88509874 | 0,88509874 | 0,88509874 | 0,88509874 | 0,88509874 | 0,88509874 | 0,88509874 |
| 20 | 1 | 0,88150808 | 0,90305206 | 0,85637343 | 0,89766607 | 0,87612208 | 0,87612208 | 0,88868941 | 0,88868941 | 0,87612208 | 0,87612208 | 0,88868941 | 0,88868941 |
| 25 | 1 | 0,88150808 | 0,96588869 | 0,91561939 | 0,91202873 | 0,91202873 | 0,91202873 | 0,91202873 | 0,91202873 | 0,91202873 | 0,91202873 | 0,91202873 | 0,91202873 |
| 30 | 1 | 0,88150808 | 0,96588869 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 |
| 35 | 1 | 0,88150808 | 0,96588869 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 |
| 40 | 1 | 0,88150808 | 0,96588869 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 | 0,9048474 |
| 45 | 1 | 0,88150808 | 0,96588869 | 0,90125673 | 0,90125673 | 0,90125673 | 0,90125673 | 0,90125673 | 0,90125673 | 0,90125673 | 0,90125673 | 0,90125673 | 0,90125673 |
| 50 | 1 | 0,88150808 | 0,96588869 | 0,91382406 | 0,91382406 | 0,91382406 | 0,91382406 | 0,91382406 | 0,91382406 | 0,91382406 | 0,91382406 | 0,91382406 | 0,91382406 |

*Tabla 82. Recall del aprobado para la asignatura de Informática en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

| Elementos por Nodo | 1 | 3 | 5 | 7 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,4640884 | 0,64088398 | 0,54972376 | 0,60220994 | 0,61878453 | 0,56629834 | 0,59944751 | 0,63812155 | 0,6519337 | 0,64364641 | 0,62430939 | 0,6519337 | 0,66298343 |
| 2 | 0,4640884 | 0,64088398 | 0,54972376 | 0,60220994 | 0,61325967 | 0,57734807 | 0,55801105 | 0,59116022 | 0,60497238 | 0,60220994 | 0,59668508 | 0,60773481 | 0,59944751 |
| 3 | 0,4640884 | 0,64088398 | 0,54972376 | 0,60497238 | 0,61878453 | 0,59116022 | 0,5718232 | 0,61049724 | 0,62154696 | 0,6160221 | 0,62983425 | 0,61049724 | 0,63259669 |
| 5 | 0,4640884 | 0,64088398 | 0,54972376 | 0,60773481 | 0,5801105 | 0,52762431 | 0,61325967 | 0,61049724 | 0,59392265 | 0,60220994 | 0,61049724 | 0,62154696 | 0,60773481 |
| 7 | 0,4640884 | 0,64088398 | 0,54972376 | 0,61325967 | 0,63535912 | 0,59116022 | 0,59944751 | 0,60773481 | 0,61049724 | 0,60773481 | 0,62707182 | 0,62154696 | 0,62154696 |
| 10 | 0,4640884 | 0,64088398 | 0,54972376 | 0,60220994 | 0,63812155 | 0,61049724 | 0,60773481 | 0,60220994 | 0,60773481 | 0,60497238 | 0,59944751 | 0,60773481 | 0,60773481 |
| 15 | 0,4640884 | 0,64088398 | 0,54972376 | 0,59392265 | 0,62154696 | 0,56906077 | 0,56353591 | 0,55524862 | 0,54696133 | 0,54696133 | 0,54696133 | 0,55524862 | 0,56353591 |
| 20 | 0,4640884 | 0,64088398 | 0,54972376 | 0,61049724 | 0,61878453 | 0,59944751 | 0,60773481 | 0,60773481 | 0,59944751 | 0,59944751 | 0,60773481 | 0,59944751 | 0,59944751 |
| 25 | 0,4640884 | 0,64088398 | 0,54972376 | 0,56906077 | 0,61325967 | 0,63812155 | 0,63535912 | 0,63535912 | 0,63535912 | 0,63535912 | 0,63535912 | 0,63535912 | 0,63535912 |
| 30 | 0,4640884 | 0,64088398 | 0,54972376 | 0,53038674 | 0,58287293 | 0,60220994 | 0,60220994 | 0,60220994 | 0,60220994 | 0,60220994 | 0,60220994 | 0,60220994 | 0,60220994 |
| 35 | 0,4640884 | 0,64088398 | 0,54972376 | 0,56629834 | 0,56353591 | 0,57734807 | 0,57734807 | 0,57734807 | 0,57734807 | 0,57734807 | 0,57734807 | 0,57734807 | 0,57734807 |
| 40 | 0,4640884 | 0,64088398 | 0,54972376 | 0,56077348 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 |
| 45 | 0,4640884 | 0,64088398 | 0,54972376 | 0,53314917 | 0,55801105 | 0,55801105 | 0,55801105 | 0,55801105 | 0,55801105 | 0,55801105 | 0,55801105 | 0,55801105 | 0,55801105 |
| 50 | 0,4640884 | 0,64088398 | 0,54972376 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 | 0,58563536 |

*Tabla 83. Recall del aprobado para la asignatura de Mecánica en función de los parámetros de profundidad y el número de elementos mínimo para formar un nodo. (Árboles de Decisión)*

**Script de la preparación y limpieza de datos**

```
ata_preins=pd.read_excel("/Users/aleja/Desktop/Datos TFG/dadespersnomespreins.xlsx")

data_faseini=pd.read_excel("/Users/aleja/Desktop/Datos TFG/qfaseini.xlsx")

data_fasenoini=pd.read_excel("/Users/aleja/Desktop/Datos TFG/qfasenoini.xlsx")

data_faseini=data_faseini[data_faseini["CODI_PROGRAMA"]==752]

data_fasenoini=data_fasenoini[data_fasenoini["CODI_PROGRAMA"]==752]

data_faseini1=data_faseini[(data_faseini["CODI_UPC_UD"]==240011)|
        (data_faseini["CODI_UPC_UD"]==240012) |
        (data_faseini["CODI_UPC_UD"]==240013) |
        (data_faseini["CODI_UPC_UD"]==240014) |
        (data_faseini["CODI_UPC_UD"]==240015) |
        (data_faseini["CODI_UPC_UD"]==240021) |
        (data_faseini["CODI_UPC_UD"]==240022) |
        (data_faseini["CODI_UPC_UD"]==240023) |
        (data_faseini["CODI_UPC_UD"]==240024) |
        (data_faseini["CODI_UPC_UD"]==240025)]


data_fasenoini1=data_fasenoini[(data_fasenoini["CODI_UPC_UD"]=="240031")|
        (data_fasenoini["CODI_UPC_UD"]=="240131") |
        (data_fasenoini["CODI_UPC_UD"]=="240132") |
        (data_fasenoini["CODI_UPC_UD"]=="240133") |
        (data_fasenoini["CODI_UPC_UD"]=="240032") |
        (data_fasenoini["CODI_UPC_UD"]=="240033")]

data=data_fasenoini1.sort_values(["CURS","QUAD","CODI_UPC_UD"]).drop_duplicates(["CODI_EXPEDIENT","CODI_UPC_UD"])

data_faseini_supera=data_faseini1.pivot_table(index="CODI_EXPEDIENT", columns="CODI_UPC_UD", values="SUPERA")

data_faseini_notes=data_faseini1.pivot_table(index="CODI_EXPEDIENT", columns="CODI_UPC_UD", values="NOTA_NUM_DEF")

lista=data_faseini_supera.columns.values.tolist()

data_faseini_intents=1/data_faseini_supera[lista]
```

```
data_faseinicial=pd.merge(left=data_faseini_notes,right=data_faseini_intents,how="left",left_on="CODI_EXPEDIENT"
,right_on="CODI_EXPEDIENT")

data_q3_supera=data.pivot_table(index="CODI_EXPEDIENT", columns="CODI_UPC_UD", values="SUPERA")

data_q3_notes=data.pivot_table(index="CODI_EXPEDIENT", columns="CODI_UPC_UD",
values="NOTA_NUM_DEF")

data_q3_curs=data.pivot_table(index="CODI_EXPEDIENT", columns="CODI_UPC_UD", values="CURS")

data_q3_linealreg=pd.merge(left=data_q3_notes,right=data_q3_curs,how="left",left_on="CODI_EXPEDIENT",right_o
n="CODI_EXPEDIENT")

data_preins=data_preins[["CODI_EXPEDIENT","NOTA_ACCES"]]

data_linealreg=pd.merge(left=data_linealreg,right=data_preins,how="left",left_on="CODI_EXPEDIENT",right_on="CO
DI_EXPEDIENT")

data_linealreg=data_linealreg.dropna()

data_linealreg.columns=["CODI_EXPEDIENT","ELECTROMAG","METODOS","MATERIALES","EDOS","INFO","ME
C","Año ELECTROMAG","Año METODOS","Año MATERIALES","Año EDOS","Año INFO","Año
MEC","ALGEBRA","CALCULO1","MECFON","QUIMICA1","FONINFO","GEOMETRIA","CALCULO2","TERMO","QUI
MICA2","EXPRE","Intentos ALGEBRA","Intentos CALCULO1","Intentos MECFON","Intentos QUIMICA1","Intentos
FONINFO","Intentos GEOMETRIA","Intentos CALCULO2","Intentos TERMO","Intentos QUIMICA2","Intentos
EXPRE","NOTA_ACCES"]

data_linealreg[['ELECTROMAG','METODOS','MATERIALES','EDOS','INFO','MEC']]=(data[['ELECTROMAG','METOD
OS','MATERIALES','EDOS','INFO','MEC']]>=5).astype(int)

data=data_linealreg
```

**Script de la Regresión Logística**

```
for e in ['ELECTROMAG','METODOS','MATERIALES','EDOS','INFO','MEC']:

    print(e)


l=['ALGEBRA','CALCULO1','MECFON','QUIMICA1','FONINFO','GEOMETRIA','CALCULO2','TERMO','QUIMICA2','EX
PRE','Intentos ALGEBRA','Intentos CALCULO1','Intentos MECFON','Intentos QUIMICA1','Intentos FONINFO','Intentos
GEOMETRIA','Intentos CALCULO2','Intentos TERMO','Intentos QUIMICA2','Intentos EXPRE','NOTA_ACCES']

    y=data[e]

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=17)

    model=LogisticRegression(C=10000000000)

    model.fit(X_train,y_train)

    print("Model Score = "  + " = "  + str(model.score(X_train,y_train)))

    predictions=model.predict(X_test)
```

```
    print( "Accuracy Score = " + str(accuracy_score(y_test, predictions)) + "\n")
    string = e + " = " + str(model.intercept_[0])
    coefs=[]
    for i in range(len(model.coef_[0])):
        string=string+ " + " + str(model.coef_[0][i]) + "*" +l[i]
        coefs.append(model.coef_[0][i])
    df["Variables"]=["Constante"]+l
    df["Coeficientes" + e]=[model.intercept_[0]] + coefs
    print(string + "\n")
    confm=confusion_matrix(y_test, predictions)
    print(confusion_matrix(y_test, predictions))
    print(classification_report(y_test, predictions))
    print("0 precision = " + str(confm[0][0]) + " / (" + str(confm[1][0]) + " + "+str(confm[0][0])+")")
    print("1 precision = " + str(confm[1][1]) + " / (" + str(confm[1][0]) + " + "+str(confm[1][1])+")")
    print("0 recall = " + str(confm[0][0]) + " / (" + str(confm[0][1]) + " + "+str(confm[0][0])+")")
    print("1 recall = " + str(confm[1][1]) + " / (" + str(confm[1][0]) + " + "+str(confm[0][0])+")")
    print("0 support = " + str(confm[0][0]) + " + "+str(confm[0][1]))
    print("1 support = " + str(confm[1][0]) + " + "+str(confm[1][1]))
```

**Script de los Árboles de Decisión**

```
df=pd.read_csv("/Users/aleja/Desktop/Datos TFG/Notas(TFG).csv")
l=[]
for a in ['ELECTROMAG','METODOS','MATERIALES','EDOS','INFO','MEC']:
    Y=df[a]
    X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=17,shuffle=True)
    l1=[]
    for i in [1,2,3,5,7,10,15,20,25,30,35,40,45,50]:
        l2=[]
        for e in [1,2,3,5,7,10,15,20,25,30,35,40,45,50]:
            tree=DecisionTreeClassifier(criterion="entropy",max_depth=i,min_samples_leaf=e)
```

```
            tree.fit(X_train,y_train)
            prediction=tree.predict(X_test)
            confm=confusion_matrix(y_test,prediction)
            print( a + " : Prof =" + str(i)+", Min Elem = " + str(e) )
            print(confm)
            l2.append(tree.score(X_test,y_test))
        l1.append(l2)
    l.append(l1)
l_asig=['ELECTROMAG','METODOS','MATERIALES','EDOS','INFO','MEC']
for e in range(len(l_asig)):
    dataframe=pd.DataFrame(
        {
            "Elementos por Nodo":[1,2,3,5,7,10,15,20,25,30,35,40,45,50],
            "1":l[e][0],
            "3":l[e][1],
            "5":l[e][2],
            "7":l[e][3],
            "10":l[e][4],
            "15":l[e][5],
            "20":l[e][6],
            "25":l[e][7],
            "30":l[e][8],
            "35":l[e][9],
            "40":l[e][10],
            "45":l[e][11],
            "50":l[e][12],

        }

    )
```

```
    dataframe.to_excel("/Users/aleja/Desktop/Datos TFG/" + l_asig[e]+"Scores_ÁrbolesDecisionBUENOS.xls")


l_0=[]
l_1=[]
for a in ['ELECTROMAG','METODOS','MATERIALES','EDOS','INFO','MEC']:
    Y=df[a]
    X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=17,shuffle=True)
    l1_0=[]
    l1_1=[]
    for i in [1,2,3,5,7,10,15,20,25,30,35,40,45,50]:
        l2_0=[]
        l2_1=[]
        for e in [1,2,3,5,7,10,15,20,25,30,35,40,45,50]:
            tree=DecisionTreeClassifier(criterion="entropy",max_depth=i,min_samples_leaf=e)
            tree.fit(X_train,y_train)
            prediction=tree.predict(X_test)
            confm=confusion_matrix(y_test,prediction)
            print( a + " : Prof =" + str(i)+", Min Elem = " + str(e) )
            print(confm)
            recall0=confm[0][0]/(confm[0][1]+confm[0][0])
            print(str(recall0))
            recall1=confm[1][1]/(confm[1][0]+confm[1][1])
            #print("0 recall = " + str(confm[0][0]) + " / (" + str(confm[0][1]) + " + "+str(confm[0][0])+")")
            #print("1 recall = " + str(confm[1][1]) + " / (" + str(confm[1][0]) + " + "+str(confm[1][1])+")")
            l2_0.append(recall0)
            l2_1.append(recall1)
        l1_0.append(l2_0)
        l1_1.append(l2_1)
    l_0.append(l1_0)
    l_1.append(l1_1)
```

```
            #print("Score testing = " + str(tree.score(X_test,y_test)))
            #print("Score para i = ",str(i)," es de ", score, "Std_deviation = ",std_score)
l_asig=['ELECTROMAG','METODOS','MATERIALES','EDOS','INFO','MEC']
for e in range(len(l_asig)):
    dataframe=pd.DataFrame(
        {
            "Elementos por Nodo":[1,2,3,5,7,10,15,20,25,30,35,40,45,50],
            "1":l_0[e][0],
            "3":l_0[e][1],
            "5":l_0[e][2],
            "7":l_0[e][3],
            "10":l_0[e][4],
            "15":l_0[e][5],
            "20":l_0[e][6],
            "25":l_0[e][7],
            "30":l_0[e][8],
            "35":l_0[e][9],
            "40":l_0[e][10],
            "45":l_0[e][11],
            "50":l_0[e][12],

        }

    )
    dataframe.to_excel("/Users/aleja/Desktop/Datos TFG/" + l_asig[e]+"Recall0.xls")
l_asig=['ELECTROMAG','METODOS','MATERIALES','EDOS','INFO','MEC']
for e in range(len(l_asig)):
    dataframe=pd.DataFrame(
        {
            "Elementos por Nodo":[1,2,3,5,7,10,15,20,25,30,35,40,45,50],
```

```
        "1":l_1[e][0],
        "3":l_1[e][1],
        "5":l_1[e][2],
        "7":l_1[e][3],
        "10":l_1[e][4],
        "15":l_1[e][5],
        "20":l_1[e][6],
        "25":l_1[e][7],
        "30":l_1[e][8],
        "35":l_1[e][9],
        "40":l_1[e][10],
        "45":l_1[e][11],
        "50":l_1[e][12],

    }

)
dataframe.to_excel("/Users/aleja/Desktop/Datos TFG/" + l_asig[e]+"Recall1.xls")
```

**Script SVM**

```
lista=['ALGEBRA','CALCULO1','MECFON','QUIMICA1','FONINFO','GEOMETRIA','CALCULO2','TERMO','QUIMICA2',
'EXPRE','Intentos ALGEBRA','Intentos CALCULO1','Intentos MECFON','Intentos QUIMICA1','Intentos
FONINFO','Intentos GEOMETRIA','Intentos CALCULO2','Intentos TERMO','Intentos QUIMICA2','Intentos
EXPRE',"NOTA_ACCES"]

for a in ['ELECTROMAG','METODOS','MATERIALES','EDOS','INFO','MEC']:
    print(a)
    target=df[a]
    X_train, X_test, target_train, target_test = train_test_split(X, target, test_size=0.3, random_state=17)
    l=[0.001,0.01,0.1,1,10,100,1000,10000,100000,1000000,10000000,100000000,1000000000]
    for i in range(len(l)):
        if i==0:
            classifier= svm.SVC(kernel="linear",C=l[i])
```

```
        classifier.fit(X_train, target_train)

        prediction=classifier.predict(X_test)

        print("C = " + str(l[i]) + ", Score = " + str(classifier.score(X_test,target_test)) + ", Number of Support Vectors =
"+str(len(classifier.support_vectors_)))

        confm=confusion_matrix(target_test,prediction)

        print(confusion_matrix(target_test,prediction))

        print(classification_report(target_test, prediction))

        print("0 precision = " + str(confm[0][0]) + " / (" + str(confm[1][0]) + " + "+str(confm[0][0])+")")

        print("1 precision = " + str(confm[1][1]) + " / (" + str(confm[1][0]) + " + "+str(confm[1][1])+")")

        print("0 recall = " + str(confm[0][0]) + " / (" + str(confm[0][1]) + " + "+str(confm[0][0])+")")

        print("1 recall = " + str(confm[1][1]) + " / (" + str(confm[1][0]) + " + "+str(confm[1][1])+")")

        print("0 support = " + str(confm[0][0]) + " + "+str(confm[0][1]))

        print("1 support = " + str(confm[1][0]) + " + "+str(confm[1][1]) + "\n")

        string="Ecuación : 0 = " + str(classifier.intercept_[0])

        for e in range(len(classifier.coef_[0])):

            string=string+ " + " + str(classifier.coef_[0][e]) + "*" +lista[e]

        print(string)

    else:

        classifier1=svm.SVC(kernel="linear",C=l[i-1])

        classifier= svm.SVC(kernel="linear",C=l[i])

        classifier.fit(X_train, target_train)

        classifier1.fit(X_train, target_train)

        prediction=classifier.predict(X_test)

        prediction1=classifier1.predict(X_test)

        if classifier.score(X_test,target_test)>classifier1.score(X_test,target_test):

            print("C = " + str(l[i]) + ", Score = " + str(classifier.score(X_test,target_test)) + ", Number of Support Vectors =
"+str(len(classifier.support_vectors_)))

            confm=confusion_matrix(target_test,prediction)

            print(confusion_matrix(target_test,prediction))

            print(classification_report(target_test, prediction))

            print("0 precision = " + str(confm[0][0]) + " / (" + str(confm[1][0]) + " + "+str(confm[0][0])+")")
```

```
print("1 precision = " + str(confm[1][1]) + " / (" + str(confm[1][0]) + " + "+str(confm[1][1])+")")
print("0 recall = " + str(confm[0][0]) + " / (" + str(confm[0][1]) + " + "+str(confm[0][0])+")")
print("1 recall = " + str(confm[1][1]) + " / (" + str(confm[1][0]) + " + "+str(confm[1][1])+")")
print("0 support = " + str(confm[0][0]) + " + "+str(confm[0][1]))
print("1 support = " + str(confm[1][0]) + " + "+str(confm[1][1]))
for e in range(len(classifier.coef_[0])):
    string=string+ " + " + str(classifier.coef_[0][e]) + "*" +lista[e]
print(string)
```